


# AN ANALYSIS OF THE IMPACT OF LOG-LINEAR REGRESSION ON THE ESTIMATED LEARNING CURVE PARAMETERS

THESIS

Tom Tracht, B.S.
Captain, USAF

AFIT/GCA/LSQ/88S-9



DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY
Wright-Patterson Air Force Base, Ohio

AFIT/GCA/LSQ/88S-9



# AN ANALYSIS OF THE IMPACT OF LOG-LINEAR REGRESSION ON THE ESTIMATED LEARNING CURVE PARAMETERS

THESIS

Tom Tracht, B.S.
Captain, USAF

AFIT/GCA/LSQ/88S-9

The contents of the document are technically accurate, and no sensitive items, detrimental ideas, or deleterious information is contained therein. Furthermore, the views expressed in the document are those of the author and do not necessarily reflect the views of the School of Systems and Logistics, the Air University, the United States Air Force, or the Department of Defense.

AFIT/GCA/LSQ/88S-9

# AN ANALYSIS OF THE IMPACT OF LOG-LINEAR REGRESSION ON THE ESTIMATED LEARNING CURVE PARAMETERS

THESIS

Presented to the Faculty of the School of Systems and Logistics
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Cost Analysis

Tom Tracht, B.S.
Captain, USAF

September 1988

## Acknowledgements

I would like to express my gratitude for thoes people who assisted in the completion of this research. Special thanks goes to my thesis advisor, Dr. Kankey, who not only got me on track but made the track easier to follow. Thanks to Rich Murphy for sharing his statistical wizardry and sense of humor. And finally, thanks to a classmate, Rod Troyanowski, who assisted me in much of the SAS programming used in this research. Needless to say this research would not be what it is without their assistance.

Along with those who directly assisted in my thesis I would like to thank those individuals who helped me maintain my sense of humor, not to mention my sanity, during this research. Thanks to all my classmates--Frank "SAC trained killer" Albanese, Crystal Blalock, Gary Boulware, Steve "are you from New Jersey" Giuliano, Linda "I feel strongly" Lyons, Pat Meyer, Tom Obringer, Chris Russick, Rod Troyanowski, and Chris Voss. This is undoubtedly the finest group I've had the priviledge to be part of.

Last, but first in my heart, I would like to thank my family who made this whole process enjoyable. To my wife, Cheri, thanks for putting up with the mess in every room of the house. To my children, Jackie, Dawna, and Katie, thanks for being angels (at least most of the time) and keeping my life in perspective. I love you.

Tom "will someone check my spelling" Tracht

## Table of Contents


| | Page |
|---|---|
| Acknowledgements | ii |
| List of Figures | v |
| List of Tables | vi |
| Abstract | viii |
| I. Introduction | 1 |
| General Issue | 1 |
| Specific Issue | 2 |
| Research Objectives | 5 |
| Definitions | 6 |
| II. Literature Review | 8 |
| History | 8 |
| Formulations | 11 |
| Cumulative Average Formulation | 11 |
| Unit Formulation | 12 |
| Characteristics of the Two Formulations | 13 |
| The Unit Formulation in Depth | 14 |
| Parameter Estimation | 14 |
| Fitting Techniques | 16 |
| Normal Equations | 17 |
| The Linear Model | 19 |
| Assumptions About the Data | 22 |
| Normality | 22 |
| Constant Variance | 27 |
| Bias | 28 |
| III. Methodology | 32 |
| Chapter Overview | 32 |
| General Method | 32 |
| Data Simulation | 32 |
| General Simulation Procedures | 33 |
| Specific Simulation Procedures | 34 |
| Fitting Techniques | 37 |
| Ordinary Least-Squares | 37 |
| Weighted Least-Squares | 38 |
| Conclusion | 39 |

Page

IV. Analysis Procedures . . . . . 40

Chapter Overview . . . . . 40
Research Objective 1 . . . . . 40
Research Objective 2 . . . . . 44
Research Objective 3 . . . . . 45
Research Objective 4 . . . . . 47
Research Objective 5 . . . . . 48
Research Objective 6 . . . . . 49

V. Findings . . . . . 51

Chapter Overview . . . . . 51
Research Objective 1 . . . . . 51
Research Objective 2 . . . . . 57
Research Objective 3 . . . . . 59
Research Objective 4 . . . . . 65
Research Objective 5 . . . . . 68
Research objective 6 . . . . . 69

VI. Conclusions and Recommendations . . . . . 73

Chapter Overview . . . . . 73
Summary of Findings . . . . . 73
Areas for Further Research . . . . . 76

Appendix A: Sample SAS Programs for Data Simulation . 78

Appendix B: Sample SAS Programs for Data Analysis . . 92

Bibliography . . . . . 98

Vita . . . . . 101

## List of Figures


| Figure | | Page |
|---|---|---|
| 1. | Illustration of a Multiplicative Error Term . . . . | 20 |
| 2. | Illustration of an Additive Error Term . . . . . | 20 |
| 3. | Rankit Plot of a Normal Distribution . . . . . | 26 |
| 4. | Rankit Plot of a Nonnormal Distribution . . . . . | 26 |

## List of Tables


| Table | | Page |
|---|---|---|
| 1. | Comparison of the Unit and Cumulative Average Curves | 11 |
| 2. | Simulation Characteristics | 34 |
| 3. | Avinger's Lot Size Characteristics | 41 |
| 4. | Lot Size Characteristics | 45 |
| 5. | Estimated First Unit Cost Using the Lot Midpoint Heuristic | 53 |
| 6. | Estimated Slope Coefficient Using the Lot Midpoint Heuristic | 53 |
| 7. | Estimated First Unit Cost Using True Lot Midpoint | 55 |
| 8. | Estimated Slope Coefficient Using True Lot Midpoint | 55 |
| 9. | Estimated First Unit Cost Using True Lot Midpoint, and Lot Size as Weights | 56 |
| 10. | Estimated Slope Coefficient Using True Lot Midpoint, and Lot Size as Weights | 57 |
| 11. | Comparison of Statistics From SAS ANOVA Tables and Hutchison's Learning Curve Program | 58 |
| 12. | Summary Table of the Individual Data Sets | 61 |
| 13. | Data Set One ( $\sigma$ = 0.04) | 63 |
| 14. | Data Set Two ( $\sigma$ = 0.12) | 63 |
| 15. | Data Set Three ( $\sigma$ = 0.20) | 63 |
| 16. | Estimated First Unit Cost Parameters with Reduced Bias | 67 |
| 17. | A Comparison, on Average, of Biased and Reduced Biased Estimates for the First Unit Cost Parameters | 68 |
| 18. | D-Statistic for $\ln(\bar{Y})$'s and $\bar{Y}$'s | 69 |
| 19. | Calculated Variances for $\ln(\bar{Y})$'s and $\bar{Y}$'s Using a Multiplicative Error Term | 70 |

Table Page

20. Calculated Variances for $\ln(\bar{Y})$'s and $\bar{Y}$'s Using an Additive Error Term . . . . . 71

21. Effect of Bias on an Estimate . . . . . 74

## Abstract

The job of estimating the cost of a Department of Defense production program often requires that the analyst use learning curve theory to estimate recurring costs. A common method for predicting these future costs is to fit the following learning curve model to production lot data from past analogous programs:

$$\bar{Y} = A * X^{**}b$$

which in it's linear form reads:

$$\ln(\bar{Y}) = \ln(A) + b*\ln(X)$$

The parameters in the equation are, X, the unit number; $\bar{Y}$, the average unit cost of the lot for unit formulation or the cumulative average cost per unit through unit X for cumulative average formulation; A, the theoretical first unit cost; and b, the slope coefficient.

This thesis involved research on the unit formulation of learning curve theory. Both ordinary and weighted least-squares fitting techniques were used. Comparisons of the ordinary and weighted least-squares techniques, true lot midpoint and a heuristic lot midpoint, and bias and unbiased results were made. Also investigated were the assumptions of normality, and constant variance of the lot data.

The data for these comparisons were simulated using the SAS software system. First, unit data was generated using

the log-linear learning curve equation shown above, using a first unit cost of 25,000 and a learning curve slope of 80 percent, and a multiplicative, normally distributed error not shown in the equation. Five hundred runs with lot size of 210 each were generated for three different error terms.

The research showed that the heuristic lot plot point biased the slope coefficient only when the weighted least-squares best fit (WLSBF) technique was used, and biased the first unit cost parameter for both ordinary and weighted least-squares. However, even when true lot midpoint was substituted for the heuristic lot plot point the first unit cost parameter remained biased, although the amount of the bias was reduced. This amount of bias present is almost solely a function of the variance in the fitted data. On the other hand, the bias in the slope coefficient was eliminated when true lot midpoint was used.

The bias in the first unit cost was on the high side, so an approximation to the bias reduction factor recommended by Ilderton was used with excellent results.

The comparison between ordinary and weighted least-squares provided mixed results when the bias was present, however, when the bias was removed the WLSBF technique was clearly superior.

Finally, the assumption of normality and constant variance of the average unit cost data was shown to be valid, although further research into both areas may be warranted.

# AN ANALYSIS OF THE IMPACT OF LOG-LINEAR REGRESSION ON THE ESTIMATED LEARNING CURVE PARAMETERS

## I. Introduction

### General Issue

All areas of the Department of Defense (DoD) are charged with the responsibility to be faithful stewards of the taxpayers money. However, in the area of weapon system procurement, the taxpayers have heard all too many reports of unfaithful stewardship. Reports of $600 dollar hammers, coffee pots costing several thousand dollars apiece, and noted scandals involving General Dynamics and General Electric have plagued the DoD (6:39). In an article entitled "Defense Procurement: A job too important for servicemen," the author claims that "cost estimates were haywire, ... the services tried to get money for new weapons by underestimating their costs..." (12:31). A small sample of such reports includes a Time magazine article which quoted a Pentagon analyst, as saying, "there is a systematic tendency to underestimate future costs" (17:12), and an article by Gansler who claims that a "typical DoD approach" is to give "unrealistically low initial estimates (so that the) program can get its nose in the budgetary tent," believing that once a program gets into the budget it will remain in the budget (13:7). The result of such reports seems to be, as stated in The AFSC Cost Estimating Handbook,

"that the entire aquisition process now revolves around and focuses on the cost (estimate) of an item" (3:1-2). Accurate cost estimating is essential to avoid these DoD "horror stories", and to allow for proper management and control decisions to be made throughout a weapon system's life cycle. But in order to obtain more accurate estimates, current techniques which have in part allowed for these DoD "horror stories", must be analysed to determine whether new techniques are needed to improve the accuracy of cost estimates. Less someday the Air Force states, depart from me ye unfaithful steward into everlasting interviews with 20/20 and 60 minutes.

## Specific Issue

In major weapon system acquisition within the DoD the use of learning curve theory in production estimates is commonplace (16:16),(25:20). Most government cost analysts are well versed in the use of the learning curve, and appreciate its importance to the production cost estimate. In fact, DoD Instructions 7000.3 dated April 1979, requires that all large acquisition programs use a learning curve in defining a Design to Unit Production Cost Value (8:6). With the Advanced Tactical Fighter (ATF) engine estimate, a one-percent change in the rate of learning would change the production estimate total by more than 1.5 billion then-year dollars. This is a change of greater than five-percent of the total production estimate.

The rate of learning choosen by the cost analyst is obviously very important, so one question is "how does the analyst decide on the appropriate rate of learning"? The two methods most often used are system analogy and contractor analogy. No matter which method is used there has to be a way to determine actual rates of learning experienced by analogous systems or selected contractors. This is where regression techniques get involved. Actual data is either ploted or entered into one of the many learning curve programs, the most popular being one developed by Ilderton, called ICLOT, and the most flexible being one developed by Hutchison. The ICLOT program uses the weighted least-squares technique, offering the user no option, while the Hutchison learning curve program offers the user the choice between ordinary and weighted least-squares to calculate the actual rate of learning (16:22), (15:vii).

As shown by Avinger either ordinary or weighted least-squares techniques provide a good fit to production lot data, but there is room for improvement in the learning curve fitting technique (5:64-70). One potential area for improvement to Avinger's thesis is to use the true lot midpoint algorithm as discussed by Hutchison, instead of the heuristic used by Avinger (5:7). Another potential area for improvement is in the weighting of the lots in order to eliminate the problem of heteroscedasticity. Currently the only weighting scheme offered by the literature is to weight each point by the number of units in the lot. This is the

correct weighting scheme if the analyst has individual unit cost data and the variance of the estimated unit cost is constant. However, since learning curve data used by the analyst is almost exclusively reported as lot data the analyst must deal with lot averages. This raises the question as to the variance of the lot average costs. This question must be answered in order to determine the proper weighting scheme. The other area for improvement is to eliminate the bias that is present as discussed by Ilderton (16:43-45) and Daneman (10) and demonstrated by Avinger (5). Each of these authors shows that a bias exist, but they do not agree on the magnitude and/or direction.

The goal of this thesis is to further the work accomplished by Avinger, who investigated various learning curve fitting techniques to determine which method most closely approximated the actual learning curve. This thesis will compare the ordinary least-squares (OLS) technique with the weighted least-squares (WLS) technique. The focus is on the WLS technique which offers the most promise for DoD production programs which are made up of unequal lot sizes, and is equivalent OLS when lot sizes are equal. The results obtained by Avinger on WLS will be compared with the results obtained through this study to determine if progress is being made towards a better fitting technique.

Research Objectives

In order to achieve these goals, the following objectives were used to guide the research:

1. Compare and contrast results from use of the true lot midpoint algorithm versus the lot midpoint heuristic for both ordinary and weighted least-squares.

2. Validate the procedures used to obtain the parameter estimates and statistics for both ordinary and weighted least-squares.

3. Compare the fitting techniques of ordinary and weighted least-squares for learning curve data. Which technique predicts closest to the true mean of the intercept and slope parameter? Which technique is a more efficient estimator of the intercept and slope parameter?

4. Determine how well the estimated value for the intercept and slope approximate the true value. Determine whether the bias reduction factor proposed by Ilderton or the one proposed by Daneman is the most appropriate for learning curve data. Upon determining which bias reduction factor is most appropriate, determine if the factor's use will result in unbiased estimates.

5. Test whether the average unit cost data is normally distributed, in both the transformed and untransformed state.

6. Test to see whether the weighting scheme used in the ICLOT program and the Hutchison learning curve program can be improved upon.

Definitions

Cost(s) - Refers to the dollar amount or hours of labor that has been or is projected to be expended on a system.

Learning Curve Theory - Refers to the regular pattern in unit cost which occurs as the contractor and contractor personnel gain experience producing a particular system. The learning curve is also referred to as the progress curve, cost improvement curve, and experience curve. The terms all mean the same thing in this thesis.

Learning Curve Slope (L) - "A nonalgebraic concept which can be visualized as the percent (or ratio) of work required to produce an item after a 100 percent increase in quantity" (19). For example, if the cost of unit four is $100, using the unit curve assumption and a learning curve slope of 80 percent, the cost of unit eight (100 percent increase from unit four) would be $100 * .80 or $80, unit 16 would cost $64, and so on. The slope is given by equation 1.1 and 1.2.

$$L = Y_{2X} / Y_X = A(2X^{**}b)/A(x^{**}b) \qquad (1.1)$$

This can be further reduced to read

$$L = 2^{**}b \qquad (1.2)$$

Which can be rewritten in terms of b, by taking the logarithm of both sides, which is the

expression used in both the cumulative average and unit formulation.

$$b = \log(L)/\log(2) \qquad (1.3)$$

Rate of Learning - Is expressed as 1 - L, when L is in decimal form.

## II. Literature Review

This literature review provides the foundation for the research of the learning curve theory as it is used within the DoD. More specifically this review will provide the tools necessary to critique and improve the most widely used learning curve formulation. This will be accomplished through review of learning curve programs and their associated fitting techniques used in the cost estimating field, both automated and heuristic. Data assumptions in order to obtain valid statistics will be reviewed, along with tests and checks of the assumptions where practical. This review will also contain a brief historical perspective of learning curve theory, and a discussion of the two most widely used learning curve formulations, namely the unit and cumulative average formulations.

### History

The initial publication on learning curves was an article entitled "Factors Affecting the Cost of Airplanes", authored by T.P. Wright in 1936. According to the article Wright began his investigation in 1922 when he started studing the variation in cost with quantity in the aircraft industry. His findings indicated that as cumulative quantity increased the labor hours per aircraft decreased in a regular pattern (30:122). This pattern when plotted on arithmetic paper is an exponentially decreasing curve, but is a straight line when plotted on log-log paper. Wright's

findings thus gave birth to the cumulative average formulation of the learning curve theory. The cumulative average formulation worked well during the 1930s and 1940s for the aircraft industry. Procurement changed after World War II. These changes included smaller quantity buys, multiple configuration changes and breaks in production. Due to these changes, the cumulative average formulation lost favor as its projections became questionable (1:1.13.4). This loss of favor stems from the fact that the cumulative average curve dampens the perturbations which occur in a program. This loss of information greatly hinders the analysts ability to estimate the resources required for future builds/buys.

In the mid 1950s the unit formulation of the learning curve was developed. In reviewing the literature it is not clear as to who should receive credit for the unit curve formulation. In Cost Improvement Analysis, a QMT180 text, credit is given to Crawford; however, in an article written by Adams credit is given to the Boeing company (2:1-2),(1:1.13.4). Who gets credit is not critical in this thesis. What matters is that "the learning curve is the most widely used tool for both estimating the cost of new DoD programs and controlling the cost of DoD programs" as stated by Lieber (20:2). And between the cumulative average and unit curve formulation the unit formulation is more widely used. As reported by Ilderton, the Defense Contract Audit Agency (DCAA) performed an agency wide survey on learning curve experiences. Of the 219 curves, which were

based on direct labor hours or cost, 93 percent of them used the unit curve formulation (16:14-16). Also the unit formulation offers more advantages to the analyst than the cumulative average formulation. Cochran reports, "the 'cumulative average formulation' is a much dampened form of the basic data, which itself may conceal important trends" (9:56). Adams supports this claim stating that "the problem with projecting from the cumulative average curve is that it does not properly reflect the drastic changes from one lot to another" (1:1.13.4). Cochran adds, "the 'cumulative average formulation' has a sharper slope in it than does the basic unit curve". This fact leads Cochran to conclude that use of the cumulative average formulation as a forcasting tool could prove disastrous to the forcaster (9:56). An even stronger comment comes from Adams who states, "it should be understood that the unit curve is mathematically correct, whereas the 'cumulative average curve' is incorrect" (1:1.13.4).

There have been several attempts to develop other formulations to the learning curve, none of which has been institutionalized within the DoD, but some deserve further investigation. These other formulations include the DeJong model, the S-model, the plateau model, and the Stanford-B model (31:304). This thesis is primarily concerned with the unit curve formulation, touching only briefly on the cumulative average formulation.

## Formulations

Presented here are the formulations for the unit curve and the cumulative average curve. However, due to the much wider acceptance of the unit curve theory only the unit curve formula will be investigated.

Cumulative Average Formulation. The cumulative average formulation, also known as the Northrup curve, is most appropriately applied to programs where initial costs are high and then settle down to a steady rate of decline. This means that if the slope and the total program cost were expected to be equal, the cumulative average formulation would have a higher initial value than the unit formulation as depicted in Table 1. How long the units remain higher is dependent upon the slope of the learning curve. Table 1 shows the different rates of decline for the two formulations given the same learning curve slope, and the same total cost.

Table 1. Comparison of the Unit and Cumulative Average Curves

| Unit # | Unit Cost-80% unit form | Unit Cost-80% cum avg |
|---|---|---|
| 1 | $100.00 | $130.52 |
| 2 | 80.00 | 78.31 |
| 3 | 70.21 | 66.08 |
| 4 | 64.00 | 59.22 |
| 5 | 59.56 | 54.58 |
| 6 | 56.17 | 51.15 |
| 7 | 53.45 | 48.47 |
| 8 | 51.20 | 46.28 |
| Total | $534.59 | $534.61 |

The basis of the cumulative average formulation is that "as the total quantity of units doubles, the (cumulative) average cost per unit decreases by some constant percentage" (2:3-3). This constant percentage decrease is refered to as the rate of learning. The formula for the cumulative average theory is:

$$Y = A * X^{**}b \qquad (2.1)$$

where:

Y = cumulative average cost of X units
A = theoretical cost of the first unit
X = cumulative unit number
b = an expression related to the rate of learning

Unit Formulation. The unit curve formulation, also known as the Boeing curve, is most appropriatly applied to programs which begin production in a relatively stable environment. This would be indicated by "hard" tooling in place, and most of the "bugs" worked out of the development system (2:2-1). Table 1 depicts how the learning progresses for the unit formulation. Note how steady the decline is when compared with the cumulative average formulation. The basis of the unit theory is that "as the total quantity of units doubles, the cost per unit decreases by some constant percentage" (2:2-1). The unit curve formula is as follows:

$$Y = A * X^{**}b \qquad (2.2)$$

where the variables are now:

Y = cost of unit X
A = theoretical cost of the first unit
X = unit number
b = an expression related to the rate of learning

Most cost data procured by the DoD reports costs accumulated by lots and not individual units. If unit costs were reported, fitting a curve to the data would be relatively problem free. This thesis deals with the problems of fitting a curve to lot data where lot averages must be used. The formula for the unit theory which deals with lot data is:

$$\bar{Y} = A * X^{**}b \qquad (2.3)$$

where the variables are now:

$\bar{Y}$ = average cost per unit in the lot
X = lot midpoint
A,b = as defined above

Characteristics of the Two Formulations. When comparing the two formulations the following characteristics, as depicted by the text Cost Improvement Analysis, should be noted:

> (1) when both are plotted on the same scale and the same basic data is used, the unit curve is lower on the scale than the cumulative average curve as long as there is learning;
> (2) when one is linear, the other is 'curvilinear' (the linear curve is the one accepted as being most appropriate);
> (3) one is most drastically 'curvilinear' only during the early units of production such as the first 20 or 30 units (refer to Table 1); and
> (4) the 'curvilinear' line tends to become a straight line and tends to parallel the other beyond approximately the 30th unit, although, theoretically, it is never quite a straight line [2:3-7].

## The Unit Curve Formula in Depth

This section will focus strictly on the unit curve formulation. Most of what is said from here on has no application to the cumulative average formulation.

Parameter Estimation. The unit curve formula has two parameters. One parameter is first unit cost represented by the letter A. The other parameter, b, is an expression related to the rate of learning. There are two common approaches used to compute these parameters. One approach is to estimate the parameters by hand. A heuristic is used to determine lot midpoint and straight forward math is used to determine average unit cost from production lot data. The lot midpoint is used as the x-value and the average unit cost is used as the y-value. These points are plotted on log-log paper. The analyst then uses a straight edge and visually best fits the data. This procedure allows the slope and first unit cost to be approximated. As Kankey explained, such graphs are advisable for data analysis, but estimation of the first unit cost and slope are not highly precise (19). If the reader desires further information on the topic of learning curves see the Cost Improvement Analysis text (2); Volumn 1, "AFSC Cost Estimating Handbook" (3:Chap 7); or Brewer's thesis (7).

The second approach is to use one of the several software programs to estimate the slope and first unit cost parameters. The two software programs reviewed are ICLOT and the Hutchison learning curve program.

ICLOT. The ICLOT learning curve program was developed by Ilderton in 1967 to calculate a learning curve slope and first unit cost based upon production lot data (16:iv).

The ICLOT program uses the weighted least-squares best fit (WLSBF) technique to fit the data. This is due to the problem of heteroscdasticity (a topic to be discussed later), which occurs when the variance of the error term is not constant. This changing variance is almost a certainty with unequal lot sizes. Ilderton states that:

> It seems reasonable to believe that the variance of the logarithm of the average labor hours required for a large lot should be less than the variance of the logarithm of the average unit labor hours required for a small lot [16:22].

Ilderton goes on to say that in order to avoid the problem of heteroscedasticity "each lot must be given a weight proportionate to the number of units in that lot" (16:22).

In the ICLOT program the user enters the production lot data into the computer, the ICLOT program then provide the following output based on the WLSBF technique:

```
Computed Value of First Unit--A     =
Regression Slope Coefficient--b     =
Improvement Curve Percentage        =
Coefficient of Correlation--R       =
Coefficient of Determination--R**2  = (16:32)
```

Hutchison Learning Curve Program. The Hutchison learning curve program was developed in 1985 by Larry Hutchison as a Masters Degree student at the Air Force

Institute of Technology (AFIT). Hutchison's objective was to develop a computer program which would provide fexibility to the user in selecting the particular learning curve application most appropriate to his needs. Hutchison succeeded in writing a program which allows the user to select either ordinary or weighted least-squares (along with other options). Within these options the user can select the unit formulation or the cumulative average formulation (15:iv).

In the Hutchison program the user enters production lot data into the computer. The program then outputs the same statistics as the ICLOT program. For a more extensive look into the options available within the Hutchison learning curve program the reader should refer to his thesis (15).

## Fitting Techniques

There are many methods which can be used to fit a curve to a given data set. Among this list are ordinary and weighted least-squares, which are parametric techniques, and median and mean slope, which are nonparametric techniques. This review is primarily concerned with the above mentioned parametric techniques since they are the ones used in the available learning curve software programs. The major difference between parametric and nonparametric techniques is that the statistics obtained in using nonparametric techniques make no assumptions about the distribution of the data in order to be valid. However, the statistics obtained in using the parametric techniques mentioned above require

the data to be normally distributed in order to be valid. If it becomes apparent in this research that the distribution of production lot data is not normal then further research into nonparametric techniques may be warrented. No applied software could be found which used a nonparametric fitting technique to provide parameter estimation to the learning curves, so it would serve no purpose to critique the technique.

Normal Equations. The learning curve equation is $\ln(Y) = \ln(A) + b*\ln(X) + \ln(E)$ where the parameters are A, the first unit cost, and b, an expression related to the learning curve slope. For an equation such as the learning curve formula the parameters can be estimated using the normal equations for the LSBF regression technique. Following are the two normal equations:

$$b = [SUM(X_i*Y_i)-(n*\bar{X}*\bar{Y})]/[SUM(X_i)-(n*\bar{X})]$$

$$A = \bar{Y} - b*\bar{X}$$

If the WLSBF regression technique is used the normal equations take on a weighting component. In the case of the two software programs reviewed this weight component is equal to the lot size. Following are the two normal equations with weights applied:

$$b = \{SUM(w_i*X_i*Y_i)-[SUM(w_i*X_i)*SUM(w_i*Y_i)]/(SUM(w_i))\} / \{SUM(w_i*X_i) -[SUM(w_i*X_i)]/SUM(w_i)\}$$

$$A = [SUM(w_i*Y_i)-b*SUM(w_i*X_i)]/SUM(w_i)$$

If all the lot sizes were equal it is clear that the

weighted normal equations would be reduced to the unweighted normal equations since the weighting component would be reduced to a constant and would divide out (26:38,167).

Since most DoD production programs, if not all, consist of unequal lot sizes, the question becomes--which technique, WLSBF or ordinary LSBF (OLSBF), is most appropriate, and if WLSBF is more appropiate what weights should be used?

Guest points to three postulates which show that WLSBF is more appropriate when the standard deviations or variance of the lot averages are not equal. First it is easy to show in theory that the variance for the lot averages is $\sigma_i^2/n$, given that $\sigma^2$ is the variance for unit i, and the data is homoscedastic (constant variance), where n is the number of units in the lot. Therefore, if the data is homoscedastic ( $\sigma_i^2 = \sigma^2$ for all i) then for different lot sizes the variance and the standard deviation of the lot averages, are not equal, which implies that the data is heteroscedastic. Guest states that "the least-squares postulate leads to the weighted mean as the best estimate" (14:18). Guest goes on to show that the minimum variance postulate, which states that "the best estimate is that which leads to the least variance $\hat{Y}$" (where $\hat{Y}$ is the predicted value of Y), is satisfied when weights are applied to bring the variance of the lot averages to equal values (14:19). In concluding, Guest shows that the maximum likelihood postulate supports the statement that the weighted mean leads to the best estimate (14:19,20). Guest also shows that the WLSBF technique leads to a more efficient estimate (14:20).

Efficiency in this case speaks to the variance about the parameter estimates. The more efficient the estimate the smaller the variance about the parameter estimates.

In this same vein Avinger conducted a study comparing four fitting techniques, two of which were OLSBF and WLSBF. Avinger's results indicated that for unequal lot sizes the unweighted mean predicted closer to the true value, and the weighted mean was a more efficient predictor (5:43). This apparent contradiction with the postulates presented by Guest will be further studied in the research portion of this thesis.

The Linear Model. The unit curve formula which handles lot data, is a multiplicative formula as indicated by equation 2.3. This formula is rewritten, and depicted as follows in equation 2.4, to include a multiplicative error term for use in regression.

$$\bar{Y} = A * X^{**}b * E \qquad (2.4)$$

The multiplicative error term produces a constant percentage error as opposed to a constant error in magnitude or an additive error. Figure 1 and 2 demonstrate the difference between a multiplicative error term and an additive error term. Note how the band about Y decreases as more units are produced in the case of the multiplicative error term (remember that the cost per unit decreases as the number of units produced increases), whereas, the band

remains parallel to Y as more units are produced with the additive error term.

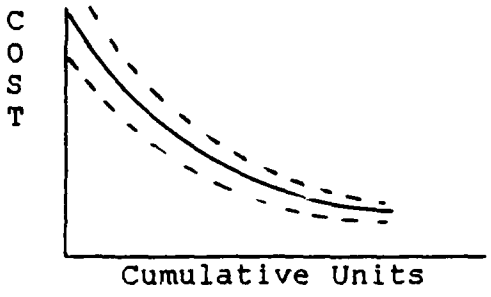


Figure 1. Illustration of a Multiplicative Error Term

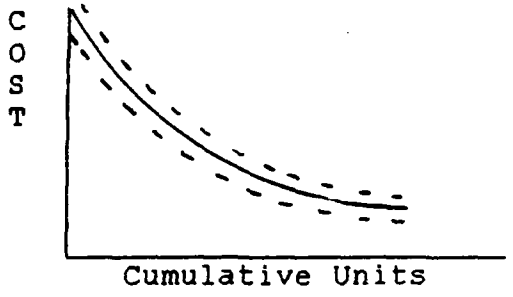


Figure 2. Illustration of an Additive Error Term

The multiplicative error term is appealing since the ability to predict the cost of a unit would seem to become more precise as the cost per unit decreased along with government and contractor experience in producing and estimating the system. The multiplicative error term is also more appealing than the additive error term since transforming equation 2.3 with an additive error term to a linear equation would be impossible.

There are some important assumptions about the E's which will follow. In order to use the linear regression, equation 2.4 must be transformed from it's curvilinear state to a linear state. This is done using a logarithmic transformation. By taking the natural logarithm (ln) of both sides of equation 2.4 the following linear equation is obtained:

$$\ln(\bar{Y}) = \ln(A) + b*\ln(X) + \ln(E) \qquad (2.5)$$

Except for the ln(E) term it is clear that equation 2.5 is

in the form of the standard linear equation $Y = a + mX$ where ln(A) is the Y-intercept (a) and b is the slope (m). This equation can now be plotted after the A and b terms are calculated, either by hand using the normal equations or by computer using a statistical software package. Of course with large data sets it becomes impractical to calculate these parameters by hand.

In order to obtain valid statistics from the LSBF technique certain assumptions about the E, or in this case the ln(E), must be met. These assumptions are normality, mean of zero, and constant variance. If these postulates are met the LSBF technique will provide the best linear unbiased estimate (18:32,171-174). These assumptions of normality and constant variance carry over to the $\ln(\bar{Y})$, but the mean is $\ln(A) + b*\ln(X)$ not zero. If the data violates any one of these, the statistics obtained from the analysis will be flawed. How badly flawed depends upon which assumption is violated and to what extreme the assumption is violated. Most of the time it is assumed that the data satisfies these postulates. The regression procedure is performed and then tests may be run on the data to test for normality, constant variance, and in the case of the error terms a pattern or trend may indicate a misspecified model or improper identification of significant variables to be included in the model (23).

## Assumptions About the Data

The assumptions about the ln(E)'s are that they are normally distributed, with a mean of zero, and a constant variance. The mean of zero assumption will not be covered in depth. Note that the least-squares technique fits the line through the data in such a way as to make the mean of the error term zero. So the assumption that the ln(E)'s have a mean of zero is automatically met when the data is fit using the least-squares technique. However, if a trend in the error terms is noted there may be a problem with the chosen model or a significant variable may have been excluded (23).

Normality. To address normality of the ln(E)'s one must also address the log-normal distribution since if the ln(E)'s are normally distributed it follows that the E's must be log-normally distributed. And if the E's are log-normally distributed it would seem reasonable that this distribution would be passed on to the $\bar{Y}$'s. People who have written on the subject, such as Avinger, have simply explained that in order for the ln(E)'s to be normally distributed the E's must be log-normally distributed (5:9). Ilderton simply states that the unit curve theory "assumes that the number of direct labor hours required to make each unit is log-normally distributed" (16:33). One might be tempted to conclude that analysts are ignoring reality to justify using the statistically powerful least-squares

technique. However, some further reading on the log-normal distribution supports the logic of the assumption.

Aitchison and Brown, in one of the few books devoted entirely to the log-normal distribution, postulate that the log-normal distribution is present in the fields of anthropology, astronomy, economics, as well as in industry. (4:100-105). The log-normal distribution is positively skewed, extending from zero to infinity (4:7-9). As Ilderton points out this is the same feasible range that program man-hours can assume. Ilderton felt that the man-hours or costs would be positively skewed since they could not take on negative values but were unbounded on the positive side (16:33). Ilderton also believed that the man-hours required to produce a unit follow the characteristic of a lognormal distribution, since as he puts it, "the variance of the man-hours required to produce a unit decreases as the number of units increases and as the expected value of the man-hours decreases" (16:33). This seems logical since the cost (or hours) required per unit is decreasing, the same magnitude error will result in a lower absolute error. This results in decreasing bounds about the estimate. Also as units are manufactured the experience gained should allow for more accurate estimates of unit cost. This also would result in a decreased variance.

This seemed to justify the initial normality assumption of the ln(E)'s until discussions on the topic with Murphy. Murphy pointed out that the assumption for the ln(E)'s was fine; however, due to the central limit theorem the lot

average values, $\ln(\bar{Y})$, did not necessarily exhibit the normal distribution of the ln(E)'s (23). The central limit theorem states:

> If random samples of n observations are drawn from a population with finite mean, u, and standard deviation, $\sigma$, then when n is large, the sample mean, $\bar{Y}$, will be approximately normally distributed with mean equal to u and standard deviation $\sigma/n$. The approximation will become more accurate as n becomes large [21:198].

What this means is that if the E's are log-normally distributed and n is small then the $\bar{Y}$'s are log-normally distributed. However if the E's are log-normally distributed and n is large then the central limit theorem indicates that the $\bar{Y}$'s are approximately normally distributed. This could have a serious impact upon the assumptions which are made about the ln(E)'s and $\bar{Y}$'s in equation 2.5. For example, if the $\bar{Y}$'s in equation 2.4 are normally distributed for large n then taking the natural logarithm of the $\bar{Y}$'s makes the $\ln(\bar{Y})$'s non-normally distributed which is a violation of the assumptions for valid statistics using LSBF.

A good question at this point is how large does n have to be for the $\bar{Y}$'s to approximate the normal distribution thus creating a violation of the normality assumption of the $\ln(\bar{Y})$'s? Mendenhall states that:

> Unfortunately there is no clear cut answer to this question, as the appropriate value for n will depend upon the populatiuon probability distribution as well as the use we will make of the approximation [21:199].

Discussions with Murphy indicate that most distributions of sample means will approximate the normal distribution when n is greater than 30; however, some distributions, such as the Cauchy distribution, do not approximate the normal distribution until n is greater than 100, while the binomial distribution approximates the normal distribution when n is no greater than 10 (23). Nonetheless the normality postulate definitely presents a problem in this research.

Although this thesis attempts to make the current fitting techniques more valid and less biased when fitting learning curve data, it is apparent that further investigation into nonparametric fitting techniques needs to be made. Nonparametric techniques make no assumptions about the distribution, thus the statistics obtained will be valid as long as the nonparametric techniques can handle data which is normal and non-normal in the same data set.

Tests For Normality. There are several methods to test for normality of the ln(E)'s and the ln(Y)'s. In practice, the ln(E)'s would not be observed, so the test for normality is done on the residuals. The residual, e, is equal to ln(Y) observed minus ln(Y) expected. Weisberg warns that if the observations or degrees of freedom for error is small the residuals may appear to be normally distributed even when they are not (29:157).

A technique for studying non-normality, espoused by Weisburg, is the normal probability plot which is often refered to as the rankit plot (29:157). Murphy explained

that the rankit plot graphs the observed distribution of either the e's or Y's versus the hypothetical or expected distribution of the e's or Y's given normality (24). Therefore, if normality were the case the plot would approximate a straight line similar to Figure 3. If normality is not the case the plot would not approximate a straight line but would appear more like Figure 4.

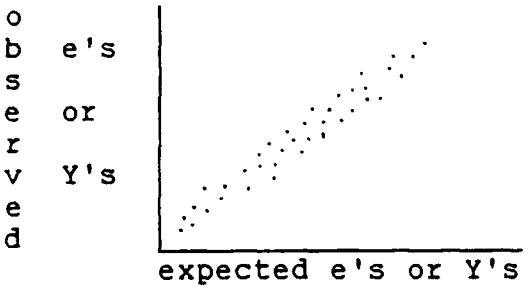


Figure 3. Rankit Plot of a Normal Distribution

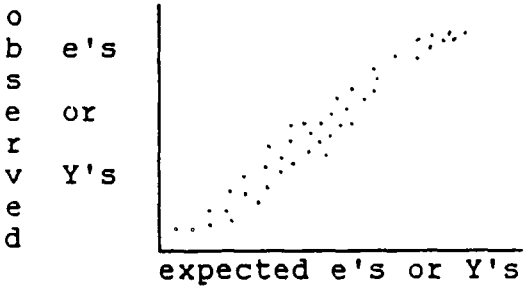


Figure 4. Rankit Plot of a Nonnormal Distribution

Although intuitive it requires practice to learn how to interprete the rankit plots (29:159).

Quantitative methods to test for normality include the Chi-squared test, the Shapiro-Wilks test, and the Kolmogorov-Smirnov (K-S) test. The Chi-squared test measures the observed frequency vesus the expected frequency based on the fact that they follow a normal distribution. This test requires a large number of observations. The Shapiro-Wilks test generates the statistic W which ranges from 0 to 1. Values of W close to one support the normality assumption while numbers close to zero lead toward rejection of the normality assumption. This test can be done with a small number of observations (24). The K-S test generates

the statistic D, which also ranges from 0 to 1. In the case of the K-S test numbers close to one lead toward rejection of the normality assumption while numbers close to zero support the normality assumption. The K-S test is the most powerful of the three test, but cannot be used when the number of observation is below fifty. The equation for D is:

$$D = \text{maximum}_i \; | F_i - S_i |$$

where:

$F_i$ = Cumulative relative frequencies of ith observation
$S_i$ = Cumulative observed frequencies of ith observation (22:458-460)

For further discussion on each of these tests and the procedures to employ them, refer to Statistics For Management And Economics, chapters 12 and 13 (22).

Constant Variance. The next requirement to use the statistical measures from the LSBF technique is constant variance of the error term over all observations. For the most part it is assumed that the variance is constant and then this assumption is tested after the modeling process has taken place. There are certain situations which definitely indicate a non-constant variance, know as heteroscedasticity. One situation is unequal lot sizes. If each observation has a variance of $\sigma^2$ then the variance for a lot of size n is $\sigma^2/n$ . If n is not equal for each lot, heteroscedasticity exist. Under normal conditions this is easily remedied by simply weighting the data by the lot size

n. This brings the variance for each lot, $\sigma^2/n$, back to $\sigma^2$ (26:170).

Individuals such as Murphy(23) and Kankey(19) have stated that weighting by the lot size may be inappropriate for learning curve data. This is due to the logarithmic transformation of the costs while the lot sizes remain untransformed, possibly giving too much weight to large lots and not enough weight to small lots. Ilderton believed that weighting by the lot size is appropriate. He states that:

> The conclusion that the weighting is appropriate follows from the fact that the variance of 'log(Y)' is approximately $(\exp^{**}\sigma^2 - 1)/n$ for large n . Furthermore ..... the value of $(\exp^{**}\sigma^2 - 1)$ is unlikely to be very different from $\sigma^2$. Consequently, combined large lots, small lots and unit data in the same analysis will not seriously affect the propriety of the weighting provided by the ICLOT progam [16:43].

Murphy points out, however, that the variance of the log(Y) is not the same thing as the variance of the $\log(\bar{Y})$ which must be considered when evaluating the unit formulation (23).

Bias

One problem found in the method for fitting a learning curve to production lot data is bias. In some cases bias is introduced in an equation in order to make the estimator more efficient. In this case however bias is introduced through the transformation process of taking the logarithm of the data and then transforming the data back to it original state. According to Daneman, the logarithm average underestimates the average of the original data and

estimates the median not the mean (10:3). Daneman believed that on the average, taking the log-linear transforms of X and Y, performing LSBF, then untransforming will produce estimates that underestimate the sample Y values (11:57). This bias will not create a large percentage error, approximately 1% to 2% according to Daneman; however, with a large production program this bias can lead to unnecessarily large dollar errors (10:5),(11:57). According to Danaman the slope parameter is accurately estimated using the LSBF regression technique while the first unit cost parameter (A) is the reason for the biased estimate (10:6). Of note is that Daneman dealt with OLSBF, non-learning curve data. Avinger's work with learning curves shows that both the slope coefficient (b) and the first unit cost may be the cause for the biased estimate (5:31-52). This research will investigate whether bias exists in the estimate of the slope. If bias does exist in the slope the literature reviewed offered no information of an adjustment factor while both Danaman and Ilderton offer bias adjustment factors for the first unit cost parameter. In order to get an unbiased or less biased estimate the first unit cost and possibly the slope parameter will have to be adjusted. Daneman suggests that:

> When we perform a log linear transform to perform least squares best fit, we then untransform to get a Y-value. To get an unbiased Y value estimate we will further multiply this estimate by 'exp**($\sigma^2$/2)' [10:7].

Ilderton states that the estimate obtained from the ICLOT program should be multiplied by:

$$\exp\{\sigma^2/2*[(-1)+(1/n)+(1/N)+[\log m_o-SUM((\log m_i)/N)]^{**}2 /n*\{SUM[(\log m_i)^{**}2]-(SUM(\log m_i))^{**}2/N\}]\}$$

where

N = number of units previously produced
n = number of units in sequence
$m_o$ = midpoint of the sequence
$m_i$ = an approximation for the lot midpoint

This equation can be segmented in four parts,

$$\exp\{\sigma^2/2*(-1)\}$$
$$\exp\{\sigma^2/2*(1/n)\}$$
$$\exp\{\sigma^2/2*(1/N)\}$$
$$\exp\{\sigma^2/2*[\log m -SUM((\log m_o)/N)]^{**}2/[n*\{SUM[(\log m_i)^{**}2 -((SUM(\log m_i))^{**}2/N]\}$$

The main influence in the equation is the first term which Ilderton shows to be a number which approximates 0.99 when $\sigma^2 = 0.1$ and 0.98 when $\sigma^2 = 0.2$. The remaining terms in the equation all approximate one, each being slightly larger than one. So the last three terms all slightly reduce the bias of the estimate. As a rule of thumb, multipling the estimated value for $\bar{Y}$ by a number slightly larger than $\exp(-\sigma^2/2)$ should closely approximate the unbiased estimate (16:43-46).

Avinger's results indicate that a number smaller than $\exp(-\sigma^2/2)$ should be used to reduce the bias in the estimate of the first unit cost or intercept term (5:36-51).

These three different conclusions will have to be investigated in this research. Avinger tends to line up with Ilderton except for the magnitude of the bias reduction

factor. This may be resolved when true lot midpoint is used to fit the data. The contradiction between Daneman and Ilderton seems more substantial, in that Daneman believed that the bias resulted in a slight underestimate of the true first unit cost while Ilderton believed that the bias resulted in a slight overestimate of the true first unit cost. This difference may be due to Daneman's focus on typical log-linear regression problems with equal weights per observation, versus Ilderton's focus on log-linear regression problems for learning curve data in lot form with lot weights. If the analyst were to select the wrong adjustment factor the result may be more than just a slight error.

# III. Methodology

## Chapter Overview

This chapter describes the methodology used to investigate the Research Objectives stated in Chapter I. It describes the development of programs to generate production lot data and how this data was used to answer the research objectives.

## General Method

Learning curve data was simulated and analyzed. The data were generated using known slope and intercept parameters, and three different error terms. The ordinary and weighted least-squares fitting techniques were used to estimate the slope and intercept parameters. Then analysis of the statistics, amount of bias present, and violations of the normality and constant variance assumptions were performed. In several of these cases comparisons were made to past research in the area of learning curve theory.

## Data Simulation

The SAS System was used to simulate learning curve data for lots from a production run. First, cost data based on the unit formulation were generated for individual units within the production run. Second, the data were grouped into unequal lots. The data simulation for each of the Research Objectives were similar; however differences were present to best meet the particular objective. These

differences and several of the SAS functions used to generate the data will be discussed in the following paragraphs.

General Simulation Procedures. The simulation of the unit cost data was accomplished using the formula $\ln(Y) = \ln(A) + b*\ln(X)$ with a SAS generated random error term. A program was writen to generate 100 production runs per data set, each production run consisting of 210 or 585 units, depending on the particular Research Objective. The exact equation used to generate the data was:

$$\ln(Y) = \ln(A) + b*\ln(X) + E \qquad (3.1)$$

where

A = 25,000 (true first unit cost)
b = ln(0.80)/ln(2) = -0.321928095 (equivalent to an 80 percent learning curve slope)
X = the sequential unit number 1-210 or 1-585
Y = cost of unit X
E = a normal random error term multiplied by the selected standard deviation

The error term E was generated using the equation:

$$E = \text{RANNOR(seed)} * \text{standard deviation} \qquad (3.2)$$

where the standard deviation was a constant which represents the mean estimating error; seed is an arbitrary number which initializes the random number generation; and, RANNOR is a SAS function which "generates an observation of a normal random variable with mean 0 and variance 1 (27:267). Notice from equation 3.1, that the data is generated in the

log-linear state. This is subsequently transformed back to the standard state.

Specific Simulation Procedures. The following table depicts the peculiarities used in simulating the unit cost data by Research Objective.

Table 2. Simulation Characteristics

| Research Objective | Error Term | # of Units Per Production Run |
|---|---|---|
| 1 and 2 | RANNOR(1446) * 0.12 | 210 |
| 3 and 4 | RANNOR(1446) * 0.04 | 210 |
| | RANNOR(1592) * 0.04 | 210 |
| | RANNOR(1958) * 0.04 | 210 |
| | RANNOR(1982) * 0.04 | 210 |
| | RANNOR(2001) * 0.04 | 210 |
| | RANNOR(1446) * 0.12 | 210 |
| | RANNOR(1592) * 0.12 | 210 |
| | RANNOR(1958) * 0.12 | 210 |
| | RANNOR(1982) * 0.12 | 210 |
| | RANNOR(2001) * 0.12 | 210 |
| | RANNOR(1446) * 0.20 | 210 |
| | RANNOR(1592) * 0.20 | 210 |
| | RANNOR(1958) * 0.20 | 210 |
| | RANNOR(1982) * 0.20 | 210 |
| | RANNOR(2001) * 0.20 | 210 |
| 5 and 6 | RANNOR(1111) * 0.12 | 585 |

The reason for using different seed numbers was to generate several unique data sets which all had the same statistical properties. Different standard deviations were used in order to determine what impact different levels of estimating error would have on the statistics obtained from the LSBF techniques. The production runs of 210 units each were selected in order to replicate Avinger's data, and the production runs of 585 units each were generated to allow

for substantially different lot sizes to test the normality and constant variance assumptions.

Lot Data. The data from each of the 100 production runs were grouped into lots. A SAS program was written which performed the grouping. The lot sizes were generated using the follow equation:

$$\text{lot size} = C_i + (S_i * \text{RANUNI(seed)}) \qquad (3.3)$$

where

$C_i$ = constant for lot i
$S_i$ = scaling factor for lot i

The RANUNI function "returns a number generated from the uniform distribution on the interval (0,1)" (27:269). For Research Objective one the identical program writen by Avinger for generating random, scaled, lot sizes was used (5:24). For research objective two, three, and four the equation remained the same, however, C and S were changed to show a different production rate scheduling profile. Production buys for these objectives begin with a small lot buy, increase to a steady production rate where lot size levels off, and then dip sharply at the end when the final units to complete production are purchased.

Peculiar lot generation techniques used for each Research Objective are covered in Chapter IV.

Lot Plot Point Generation. For research objective one both the heuristic lot plot point and the true lot midpoint were calculated. The heuristic lot plot point is simply the number of units in the lot divided by two added

to the previous number of units produced. For example, if 10 units had been previously produced, and a lot plot point is to be calculated for the next lot which consists of 30 units, the lot plot point is 25 (10 + (30/2)). The one exception is with the first lot. If the first lot is equal to or greater than 10 units the lot plot point is calculated by the number of units in the lot divided by three. The true lot midpoint requires more calculations, and more extensive programing. Since the true slope is known (80 percent) the task is greatly simplified. The equation for true lot midpoint is:

$$X\text{-mid} = \{[L_i^{**}(b+1) - (F_i - 1)^{**}(b+1)] / [n_i * (b+1)]\}^{**}(1/b) \qquad (3.4)$$

where

X-mid = true lot midpoint
$L_i$ = last unit in lot i
b = expression related to the rate of learning
$F_i$ = first unit in lot i
$n_i$ = number of units in lot i (15:33-41)

For detail on how to find the true lot midpoint when the slope is not known the reader is refered to Chapter III of Hutchison's thesis (15).

The following procedures were used to include the calculation of the true lot midpoint into the SAS program used to generate lot data.

1) Initialize the vector element in the SAS program to zero.
2) Calculate the lot size (Z).
3) Initiate a loop for I equaling one to Z.

4) Within the loop calculate a variable (DUM) which equals the previous value of that variable (DUM) added to the sequential unit number raised to the power of b.

5) End the loop.

6) Calculate equation 3.4.

For lot four (I=4) these procedures appear as follows within the SAS program.

```
B = LOG(.8)/LOG(2);
LOT[4,4] = 0;
Z = LOT[4,1] - LOT[3,1];          {yields the lot size}
Do I = 1 to Z;
  DUM = LOT[4,4] + ((I + LOT[3,1])**B);
  LOT[4,4] = DUM;
END;
LOT[4,4] = (LOT[4,4]/Z)**(1/B);
```

The position in the matrix designated Lot[4,4] will then be the value of the true lot midpoint for lot four for each of the 100 production runs per data set.

## Fitting Techniques

This section will cover the programs used to calculate the learning curve parameters using both the ordinary and weighted least-squares techniques.

Ordinary Least-Squares. The OLSBF technique was run on the SAS system. The data in its untransformed state is curvilinear which makes it unsuited for linear regression techniques such as least-squares. The data to be fitted, lot average cost ($\bar{Y}$) and lot midpoint (LMP), were

transformed as follows:

$$y = \ln(\bar{Y})$$

$$x = \ln(LMP)$$

In SAS the natural logarithm function is LOG as opposed to ln as used in this thesis (27:232). These points were then fit using the following SAS statements:

```
PROC REG;
MODEL y = x;
```

The PROC REG statement along with the MODEL statement will fit a line through the x,y plot point which minimizes the sum of squared errors, placing equal weight on each x,y plot point. The PROC REG and MODEL statement are defined and described in SAS User's Guide: Statistics (28:658-661).

Weighted Least-Squares. The same technique as described above is used for the WLSBF technique except for the addition of a weighting component. The selected weighting scheme for this research was lot size (Z). The same transformation as described above is made, however, the weights remain untranformed. These points were then fit using the following SAS statements:

```
W = Z;
PROC REG;
MODEL y = x;
WEIGHT BY W;
```

The PROC REG and MODEL statments fit a line through the x,y plot points which minimizes the sum of squared errors,

however, the x,y plot points are given different levels of attention depending on the lot size (weight) for that particular x,y plot point. In essence the weight creates the number of x,y plot points equal to the size of the weight. So an x,y plot point with a lot size of five will be given one-forth the attention of an x,y plot point with a lot size of twenty. For a further description on the WEIGHT statement refer to SAS User's Guide: Statistics (28:662).

## Conclusion

The described data simulation, lot data generation, true lot midpoint calculation, and fitting techniques were the steps necessary to begin the analysis which is contained in Chapter V. Chapter IV which follows, is a detailed step by step description of the analysis procedures, and is intended for the reader who is interested in precisely how the data presented in Chapter V was generated, and/or the reader who desires to replicate the research.

# IV. Analysis Procedures

## Chapter Overview

This chapter describes the steps used in the analysis of the unit formulation of the learning curve. The intent of this chapter is to provide detailed information of the procedures used in this research to allow thorough analysis and/or replication of the research. Thus, a step by step description of the procedures used for each research objective are recorded. Not recorded are the steps taken to simulate the learning curve data, to group the data into lots, and to fit the data using both ordinary and weighted least-squares. These steps are recorded in Chapter III.

Each objective is restated below, followed by the means of researching it, the SAS procedures used, and the statistical methods of analysis used.

## Research Objective 1

> Compare and contrast results from use of the true lot midpoint versus the lot midpoint heuristic for both ordinary and weighted least-squares.

STEP 1 - The first order of business was to replicate the learning curve data which was simulated by Avinger. The learning curve data was simulated using equation 3.1. Program I of Appendix A, shows the SAS program used in the simulation process. The program generated a data file which included a column for unit number, unit cost, and cumulative cost. These columns each contained 100 production runs of 210 units each, or 21,000 data points per column.

STEP 2 - A SAS file was written which would create a 7x5 matrix based on the data from step one. The columns included cumulative units, cumulative cost, lot cost, heuristic lot plot point, and average unit cost for the lot. Information for each lot was on a separate row. To handle the 100 production runs 100 such matrices were generated. Program II of Appendix A, shows the SAS program used to accomplish this step. The resulting data file was made up of five columns which included cumulative lot size, cumulative cost, lot cost, heuristic lot plot point, and average unit cost for the lot.

Generation of unequal lot sizes was accomplished using a form of equation 3.3. The seed number used was 1515, and the scaling factor and constant (actually a set range) were as shown in Table 3, portions of which were extracted from Avinger's thesis (5:24).

Table 3. Avinger's Lot Size Characteristics

| Lot Number | Smallest Lot Size | Largest Lot Size | Scaling Factor |
|---|---|---|---|
| 1 | 2 | 10 | 10 |
| 2 | 15 | 25 | 100 |
| 3 | 20 | 30 | 100 |
| 4 | 25 | 35 | 100 |
| 5 | 30 | 40 | 100 |
| 6 | 40 | 50 | 100 |
| 7 | 20 | 78 | N/A |

The program was written so that lot sizes would be generated until they fell within the prespecified ranges. Lot seven was simply 210 minus the cumulative total through lot six.

STEP 3 - The data file generated in step two, in particular the average unit cost and the heuristic lot plot point variables, were used in the SAS program shown in Program I and II of Appendix B. The first program was used for the execution of the LSBF technique. The second program was used for the execution of the WLSBF technique. A peculiarity was noted in Avinger's weighting scheme. In his thesis Avinger clearly indicates that he intended to weight by lot size (5:25); however, the SAS program which he wrote and the results he obtained indicate that he weighted by cumulative lot size. This code will be altered for the remaining research objectives; however, this research objective was to make a comparison of Avinger's data when true lot midpoint was substituted for the lot plot point heuristic. This step generated a list file which was made up of 100 ANOVA tables. Each ANOVA table has a predicted first unit cost (the term first unit cost is used interchangeably with intercept) and slope coefficient parameter for each production run.

STEP 4 - The predicted intercept and slope coefficient parameters were compiled and averaged using the PROC MEANS procedure (27:960). The intercept and slope parameters are returned to their standard state using the antilogarithm function. Mean values are then calculated. These values were then compared with Avinger's results, as reported in his thesis to ensure that the data was precisely replicated (5:43,44). The SAS program used to execute this step is Program III of Appendix B.

STEP 5 - The permanent SAS data file containing the predicted intercept and slope coefficient parameters from Step 3 were used in Program IV of Appendix B. Using the PROC UNIVARIATE PLOT NORMAL command, data was generated which included the geometric mean of the intercept parameter, the range of the parameters, and maximum and minimum values of the parameters. The data were tabled and included in this research.

STEP 6 - The procedure discussed in Chapter III for substituting true lot midpoint for the lot plot point heuristic was accomplished. At this point the matrix spoken of in Step 2 was increased to a 7x6 matrix with the additional column being lot size. The resulting program is shown in Program III of Appendix A.

STEP 7 - Step 3, 4 and 5 were repeated except this time true lot midpoint data was used.

STEP 8 - The results obtained from Step 7 were compared with the results obtained from Steps 4 and 5. In particular the mean predicted first unit cost and slope coefficient were compared to see which resulted in a closer prediction to the true value and which was a more efficient predictor.

STEP 9 - The weighting scheme was changed from cumulative lot size to lot size. The program is shown in Program V of Appendix B. Only the results for the WLSBF technique changed. This program was executed, Steps 4 and 5 were repeated, and the results were once again compared. The results obtained from this step would provide the means

for validating the procedures and programs accomplished to this point.

Research Objective 2

Validate the procedures used to obtain the parameter estimates and the statistics for both ordinary and weighted least-squares

Step 1 - Three ANOVA tables from both ordinary and weighted least-squares were selected. The production runs were 3, 33, and 82.

STEP 2 - The first unit in the lot, the last unit in the lot, and the total lot cost were input into a data file in the Hutchison learning curve program. With these inputs the software calculated the true lot midpoint using an iterative process, and calculated the average unit cost for the lot (15:26-68). This was the same data which were fit by the least-squares technique in this research.

STEP 3 - The statistics obtained from the Hutchison learning curve program were compared with those statistic contained in the ANOVA table. The statistics of concern were the slope coefficient, learning curve slope, first unit cost, and coefficient of correlation (R) and determination (R**2).

Validation of the procedures and program used to this point would be based on whether the statistics were approximately equal, allowing for rounding error and necessary procedural differences.

For Research Objectives 3 and 4, the simulated data goes from one data set to fifteen. Each data set is unique,

either using a different error term or seed number. The data is fitted using both ordinary and weighted least-squares. Table 2 in Chapter III should be refered to for peculiarities in the simulation of each data set.

Research Objective 3

> Comparing the fitting techniques of ordinary and weighted least-squares for learning curve data, which technique is better at predicting the true intercept and slope, and which technique is a more efficient estimator of the intercept and slope.

STEP 1 - The method used to generate lot sizes in Step 2 of Research Objective 1 was modified. The changes were made to C and S of equation 3.3 while the seed number remained 1515. Table 4 shows the details:

Table 4. Lot Size Characteristics

| Lot Number | C | S |
|---|---|---|
| 1 | 5 | 5 |
| 2 | 15 | 5 |
| 3 | 25 | 10 |
| 4 | 40 | 10 |
| 5 | 40 | 10 |
| 6 | 40 | 10 |
| 7 | N/A | N/A |

Program IV of Appendix A shows the programming code used in the simulation of the lot data for this Research Objective, and Research Objective 4. This placed much tighter controls on the lot sizes while still allowing some variation. Lot seven was the remaining number of units required to bring the total to 210 units.

STEP 2 - Step 7 of Research Objective 1 was repeated for all data sets.

STEP 3 - A table was set up as follows and the data obtained from Programs III and IV of Appendix B were used to fill it in.

| Fitting Technique / Parameter | Est Intercept Parameter | Est Slope Parameter | Range of Intercept Parameter | Range of Slope |
|---|---|---|---|---|
| OLSBF | | | | |
| data set 1 | | | | |
| 2 | | | | |
| : | | | | |
| : | | | | |
| 15 | | | | |
| WLSBF | | | | |
| data set 1 | | | | |
| 2 | | | | |
| : | | | | |
| : | | | | |
| 15 | | | | |

The data from each of the identical data sets were compared to determine which technique predicted closer to the true mean and which technique was the more efficient estimator.

STEP 4 - The data sets were sorted by equal error terms and combined. Another comparison was made of the techniques. The belief was that due to the large number of observations per data set (105,000 units combined into 3,500 lots) the impact of a rare event would be greatly watered down. Therefore, in order to choose one technique over the other, that technique would have to be shown best for each of the three data sets otherwise no conclusion would be drawn in this particular objective.

This Research Objective may at first appear to be a waste of time since the next Research Objective also compared the least-squares techniques after the bias is reduced. However, current programs do not reduce the bias, but some do offer an alternative between ordinary and weighted least-squares (e.g. Hutchison's learning curve program). This objective then deals with the realistic choices available today--should the analyst use ordinary or weighted least-squares when the option is available.

Research Objective 4

> How well does the estimated value for the intercept and slope approximate the true value. Determine whether the bias reduction factor proposed by Ilderton or Daneman is the most appropriate for learning curve data. Upon determining which factor is most appropriate, determine how well the bias reduction factor does in bringing the estimated value of the intercept to the true value of the intercept.

STEP 1 - Using the data generated in Steps 2 and 3 of Research Objective 3, a comparison was made of the predicted first unit cost against the true first unit cost value of 25,000, and of the predicted slope coefficient against the true slope coefficient value of -0.321928095. These comparisons were made by individual data sets, and by grouping the data set by common error terms.

STEP 2 - Upon comparing the estimated values of the slope and intercept, determining which bias adjustment factor, if either, was most appropriate for learning curve data was simple a matter of seeing whether the bias was on the low side or the high side. Daneman's recommended bias adjustment factor adjusts the estimate upwards while

Ilderton's recommended adjustment factor adjusts the estimate downward.

STEP 3 - Upon adjusting the intercept and slope parameter as required Step 1 of this Research Objective was repeated, only this time using the adjusted data.

Research Objectives 5 and 6 use a different simulated data set as indicated in Step 1 of Research Objective 5. The areas of concern were the normality and constant variance assumption of the $\ln(\bar{Y})$. Comparisons in these two objectives would be made between lots, not data sets. Thus, a significant difference in lot sizes were desired.

Research Objective 5

Test whether the average unit cost data is normally distributed, in both the transformed and untransformed state.

STEP 1 - The simulated data for 585 units was combined into four lots. This required modifications to programs I and IV of Appendix A. Programs V and VI of Appendix A show these modified programs. The lot sizes were set, where the first lot contained one unit, the second lot contained eight units, the third lot contained sixty-four units, and the fourth lot contained five hundred and twelve units. The data set generated by Program VI of Appendix A, was made up of 100 4x6 matrices, where the columns were the same as described in Step 2 of Research Objective 1 with the addition of lot size.

STEP 2 - A SAS program was written (see Program VI of Appendix B) which in part created four separate data sets.

The basis for the grouping was lot size, where the first data set included 100 $\bar{Y}$'s of lot size one, the second data set included 100 $\bar{Y}$'s of lot size eight, and so on. The $\bar{Y}$'s were also transformed using the natural logarithm function since the distributions of the $\ln(\bar{Y})$'s are of equal concern with the $\bar{Y}$'s. After the data was grouped and transformed the SAS statements PROC UNIVARIATE PLOT NORMAL were used (27:1182). The UNIVARIATE statement is used to "provide information on the distribution of a variable" (27:1181). The PLOT statement causes PROC UNIVARIATE to generate several plots including the rankit plot which was discussed in the literature review while the NORMAL statement causes PROC UNIVARIATE to calculate the Kolomogorov-Smirnov D-statistic again discussed in the literature review (27:1182).

STEP 3 - The D-statistics obtained from Step 2 were tabled and compared to the selected critical value. The hypothesis was that the $\ln(\bar{Y})$'s were normally distributed. The selected critical value was 0.122 which is at the 10 percent significance level. If the computed D-statistic is greater then the critical value the hypothesis will be rejected at the 90 percent confidence level.

Research Objective 6

> Test to see whether the weighting scheme used in the ICLOT program and the Hutchison learning curve program can be improved upon.

STEP 1 - The same data generated in Step 1 of Research Objective 5 was used.

STEP 2 - The data obtained from Step 1 were sorted by lot size, thus creating four sub-data sets. The $\bar{Y}$'s were then transformed using the natural logarithm function. The SAS statements PROC MEANS VAR were used, with the $\bar{Y}$'s and the $\ln(\bar{Y})$'s as the specified variables. This program would calculate the variance of both the $\bar{Y}$'s and the $\ln(\bar{Y})$'s. This program is contained in Appendix B as Program VII.

STEP 3 - The variance of the $\bar{Y}$'s and the $\ln(\bar{Y})$'s for each data set were compared. If weighting by lot size is correct the variance of the $\ln(\bar{Y})$'s for each of the data sets will differ by the inverse of the lot size. For example, the variance of the $\ln(\bar{Y})$'s for lots of size eight should be one-eighth as large as the variance of the $\ln(\bar{Y})$'s for lot sizes of one.

STEP 4 - Steps one through three were repeated with the only change being to the equation used to generate the data (see Program VII of Appendix A). The following equation was used:

$$Y = A * X^{**}b + E$$

where

E = RANNOR(1111) * 500
A,X,b as previously defined

The difference in this equation is the error term. The previously generated data used a multiplicative error term where the data is now being generated with an additive error term. The reason for this change will be discussed in the results section for Research Objective 6.

# V. Findings

## Chapter Overview

This chapter presents the results of the research objectives. There were a number of interesting findings both confirming and refuting some past beliefs. The findings will be presented by research objectives. Each Research Objective will be restated, followed by the results of the research.

## Research Objective 1

> Compare and contrast results from use of the true lot midpoint algorithm versus the lot midpoint heuristic for both ordinary and weighted least-squares.

The first step was to replicate Avinger's results for both ordinary and weighted least-squares. This provided a baseline on which improvements and findings could be made. The changes to Avinger's methodology were made and documented step by step in order to see the significance of each change.

After going through many of Avinger's SAS files, copying them onto the system, and executing the programs, the replication effort was successfully accomplished. Table 5 shows the result obtained for the estimate of the first unit cost parameter which can be compared to the results in Avinger's thesis (5:43). The one addition to this table is the inclusion of the geometric mean (GM), which is the result of taking the average of the logarithms, for the estimated first unit cost parameter. The arithmetic

mean (AM) is the value Avinger reports as the mean. As can be seen from the the table the geometric mean is slightly below the arithmetic mean. This relationship always holds true, and is most likely the basis for Daneman's adjustment factor in multiplicative regression models. This mathmatical property can best be shown through the following example.

| X | ln(X) |
|---|---|
| 23,000 | 10.04325 |
| 24,000 | 10.08581 |
| 25,000 | 10.12663 |
| 26,000 | 10.16585 |
| 27,000 | 10.20359 |

The average of the X's ($\bar{X}$) is clearly 25,000, and the average of the ln(X)'s ($\overline{\ln(X)}$) is 10.125026. If $\overline{\ln(X)}$ is transformed, using the antilogarithm function, it might be expected to equal $\bar{X}$ or 25,000. However, since logarithms do not preserve averages $\exp(\overline{\ln(X)})$ is equal to the geometric mean of 24,960 (exp(10.125026)).

ICLOT and the Hutchison learning curve program both yield the arithmetic mean, since neither program deals with logarithmic averages in the same fashion depicted in the example. So the focus will be on the arithmetic mean of the estimated parameter.

Table 6 shows the results obtained for the estimate of the slope coefficient, b. Unlike the first unit cost, the slope coefficient is not calculated by taking the average of the logarithms, so only the arithmetic mean is reported.

These results are compared to the results obtained in Avinger's thesis (5:44).

Table 5. Estimated First Unit Cost
Using the Lot Midpoint Heuristic

True First Unit Cost = 25,000

| | OLSBF Technique | WLSBF Technique |
|---|---|---|
| Maximum | 29,167 | 28,825 |
| Mean: | | |
| GM | 25,155 | 25,755 |
| AM | 25,222 | 25,788 |
| Minimum | 20,991 | 22,121 |
| Range: | | |
| Total | 8,177 | 6,704 |
| 1st-3rd Quartile | 2,587 | 1,751 |
| Bias: | | |
| GM Mean | 155 | 755 |
| AM Mean | 222 | 788 |

Table 6. Estimated Slope Coefficient
Using the Lot Midpoint Heuristic

True Slope Coefficient = -.321928095

| | OLSBF Technique | WLSBF Technique |
|---|---|---|
| Maximum | -0.279564 | -0.293522 |
| Mean | -0.321114 | -0.326685 |
| Minimum | -0.353979 | -0.350645 |
| Range: | | |
| Total | 0.074415 | 0.057123 |
| 1st-3rd Quartiles | 0.023640 | 0.014952 |
| Bias: | | |
| Mean | 0.000814 | 0.004757 |

The reason for recording the 1st-3rd quartile range is to eliminate those outliers which may make the range for a particular technique appear worse or better than it really is. The bias is simply the predicted parameter's mean minus

the true parameter's value. As seen from Avinger's results the OLSBF technique predicted closer to the true first unit cost while the WLSBF technique was the more efficient estimator of the true first unit cost. Both ordinary and weighted least-squares techniques overestimated the true value.

Avinger's results indicate that the OLSBF technique predicted closer to the true slope coefficient while the WLSBF technique was the more efficient estimator. The OLSBF parameter estimate equates to an 80.045 percent learning curve slope, and the WLSBF parameter estimate equates to a 79.736 percent learning curve slope. Note that the bias of the WLSBF estimated slope coefficient is nearly six times that of the bias for OLSBF. This result is surprising since the only apparent difference between the two techniques was the weighting scheme.

The fact that the OLSBF technique is the best predictor of the true first unit cost and slope coefficient goes against what is taught in statistical textbooks as mentioned in the Literture Review section of this thesis. Much of this research is to determine whether these results are due to some anomaly of this data, learning curve data in general, or due to the methods used for fitting.

The first modification to the methodology was to incorporate true lot midpoint into a SAS program. Incorporating true lot midpoint into the program changed the results significantly as can be seen in Tables 7 and 8. Apparently bias still exists in the estimate of the first

unit cost parameter although the bias has been substantially reduced from 222 to 105 for OLSBF, and from 788 to 170 for WLSBF. The slope coefficient parameter is extremely close to the true value which is a marked change for the WLSBF technique. The bias in the WLSBF estimate of the slope coefficient parameter was reduced by a factor of greater than 200, and by more than 20 percent for OLSBF.

Table 7. Estimated First Unit Cost Using True Lot Midpoint

True First Unit Cost = 25,000

| | OLSBF Technique | WLSBF Technique |
|---|---|---|
| Maximum | 28,507 | 28,498 |
| Mean: | | |
| GM | 25,069 | 25,145 |
| AM | 25,105 | 25,170 |
| Minimum | 21,567 | 21,814 |
| Range: | | |
| Total | 6,940 | 6,684 |
| 1st-3rd Quartiles | 2,095 | 1,517 |
| Bias: | | |
| GM Mean | 69 | 145 |
| AM Mean | 105 | 170 |

Table 8. Estimated Slope Coefficient Using True Lot Midpoint

True Slope Coefficient = -.321928095

| | OLSBF Technique | WLSBF Technique |
|---|---|---|
| Maximum | -0.287046 | -0.290792 |
| Mean | -0.321270 | -0.321950 |
| Minimum | -0.348956 | -0.349944 |
| Range: | | |
| Total | 0.061910 | 0.059151 |
| 1st-3rd Quartiles | 0.016381 | 0.014306 |
| Bias: | | |
| Mean | 0.000658 | 0.000022 |

These changes, as noted in Tables 7 and 8, were solely attributable to changing from the lot plot poin heuristic to true lot midpoint. Even though it appears that bias is not present in the slope coefficient parameter when true lot midpoint is incorporated, no conclusion was to be drawn until Research Objective 3 had been accomplished.

While investigating the program's weighting scheme it was noted that lot plot points were weighted by cumulative units instead of the more conventional weighting, which by lot size. Avinger's intent had been to weight by lot size, as indicated in his thesis (5:25). The figures in the previous four tables were all generated using cumulative lot size as the weights in order to demonstrate the impact of going from the lot midpoint heuristic to the true lot midpoint. Tables 9 and 10 show how the figures change as the weighting scheme is changed.

Table 9. Estimated First Unit Cost Using True Lot Midpoint, and Lot Size as Weights

True First Unit Cost = 25,000

| | WLSBF Technique |
|---|---|
| Maximum | 28,291 |
| Mean: | |
| GM | 25,125 |
| AM | 25,144 |
| Minimum | 22,465 |
| Range: | |
| Total | 5,826 |
| 1st-3rd Quartile | 1,439 |
| Bias: | |
| GM Mean | 125 |
| AM Mean | 144 |

Since the weighting scheme for OLSBF remains unchanged, the values in Tables 7 and 8 do not apply in Tables 9 and 10.

Table 10. Estimated Slope Coefficient Using True Lot Midpoint, and Lot Size as Weights

True Slope Coefficient = -0.321928095

| | WLSBF Technique |
|---|---|
| Maximum | -0.296578 |
| Mean | -0.321774 |
| Minimum | -0.348465 |
| Range: | |
| Total | 0.051886 |
| 1st-3rd Quartile | 0.012480 |
| Bias: | |
| Mean | 0.000154 |

Weighting by lot size versus cumulative lot size improved all the values except for the arithmetic and geometric mean of the predicted slope coefficient. The predicted first unit cost parameter is more than one-half a percent from the true first unit cost while the predicted slope coefficient is less than one-twentieth of a percent from it's true value.

Research Objective 2

> Validate the procedures used to obtain the parameter estimates and statistics for both ordinary and weighted least-squares.

This objective was accomplished to insure that the processes used in this research, in particular the weighting scheme and the true lot midpoint algorithm, were the same or close approximations to the processes used in Hutchison's learning curve program. Table 11 shows how the relavent

statistics compared for the three production runs chosen from data generated in Program V of Appendix B.

Table 11. Comparison of Statistics From SAS ANOVA Tables and Hutchison's Learning Curve Program

| | Hutchison | ANOVA |
|---|---|---|
| OLSBF: | | |
| Production Run 3: | | |
| First Unit Cost | 24,075.87 | 24,067.38 |
| Slope Coefficient | -0.31129 | -0.31123 |
| Slope | 80.59 | 80.59 |
| R | -0.99864 | -0.99864 |
| R**2 | 0.99728 | 0.99728 |
| Production Run 33: | | |
| First Unit Cost | 26,203.10 | 26,215.90 |
| Slope Coefficient | -0.33254 | -0.33264 |
| Slope | 79.41 | 79.41 |
| R | -0.99902 | -0.99902 |
| R**2 | 0.99804 | 0.99804 |
| Production Run 82: | | |
| First Unit Cost | 23,846.22 | 23,827.22 |
| Slope Coefficient | -0.30917 | -0.30901 |
| Slope | 80.71 | 80.72 |
| R | -0.99935 | -0.99935 |
| R**2 | 0.99869 | 0.99869 |
| WLSBF: | | |
| Production Run 3: | | |
| First Unit Cost | 24,915.16 | 24,912.00 |
| Slope Coefficient | -0.31949 | -0.31946 |
| Slope | 80.14 | 80.14 |
| R | -0.99775 | -0.99775 |
| R**2 | 0.99550 | 0.99550 |
| Production Run 33: | | |
| First Unit Cost | 27,001.46 | 27,026.66 |
| Slope Coefficient | -0.33935 | -0.33954 |
| Slope | 79.04 | 79.03 |
| R | -0.99837 | -0.99837 |
| R**2 | 0.99674 | 0.99674 |
| Production Run 82: | | |
| First Unit Cost | 24,563.20 | 24,554.55 |
| Slope Coefficient | -0.31576 | -0.31569 |
| Slope | 80.34 | 80.35 |
| R | -0.99900 | -0.99900 |
| R**2 | 0.99800 | 0.99800 |

The immediate observation is that the first unit cost estimate, slope coefficient and thus the slope, are not

exactly equal for the outputs from the Hutchison learning curve program and those from the corresponding ANOVA table. However, the coefficient of correlation and thus the coefficient of determination are identical for the two outputs. The small differences in the first unit cost and slope coefficient can be explained. First, there is the obvious possibility of some rounding error, since Hutchison's learning curve program was executed on a micro computer, and SAS was run on a VAX 11/785 mainframe. Second, determination of true lot midpoint for Hutchison's program involves an iterative process where the slope used to calculate the true lot midpoint is closely approximated (15:33-42). This slope is thus best fit to the sample. In this research a known slope of 80 percent was used in the calculation of the true lot midpoint. If the simulated data had a slope different from 80 percent the the calculation of the true lot midpoint, which was one of the variables regressed, would be slightly off and different from the Hutchison's calculated true lot midpoint. In the SAS PROC REG procedure the accuracy of the true lot midpoint will affect the estimate of the first unit cost parameter.

Research Objective 3

> Compare the fitting techniques of ordinary and weighted least-squares for learning curve data. Which technique predicts closer to the true intercept and slope parameter? Which technique is the more efficient estimator of the intercept and slope parameter?

At this point the research expands from one data set to fifteen. The data sets were considered separately and in

combination. The rationale used in combining the data sets were standard deviation, either 0.04, 0.12, or 0.20, used in the generation of the data. So other than using a different seed number, the combined data were generated using identical first unit cost, slope, and standard deviation. Since all the data sets from here on use true lot midpoints, and lot size as the weights, the tables will no longer identify these facts. Also, since there would be limited value in building complete tables for the individual data sets, only a summary table was built. Refer back to Table 2 of the methodology section for specific information on how each data set was generated.

Table 12 shows the total range for both parameters in each data set, and the estimated values for the first unit cost and slope coefficient parameters. Next to the parameter is a plus or minus sign. The plus sign indicates that the estimate was greater than the true value, and the minus sign the converse. If bias is not present in the parameter estimates there should be a good balance between plus and minus signs.

Comparing OLSBF against WLSBF for accuracy of prediction showed that the OLSBF technique predicted closer to the true first unit cost in nine out of fifteen cases. For predicting the true slope coefficient the WLSBF technique predicted closer in eleven of the fifteen cases. In comparing the efficiency of the two techniques based on range, WLSBF had a smaller range, and thus was more efficient in twelve of the fifteen cases when predicting the

first unit cost and slope coefficient parameters. In the three cases where OLSBF was more efficient than WLSBF, based on total range, the common feature was the seed number used in the generation of the random error term (seed = 1958).

Table 12. Summary Table of the Individual Data Sets

| Fitting Technique/ | First Unit Cost Parameter | | Slope Coefficient Parameter | |
|---|---|---|---|---|
| | Estimate | Range | Estimate | Range |
| OLSBF | | | | |
| Data Set 1 | 24,990- | 1,891 | -0.32171+ | 0.01922 |
| 2 | 24,981- | 2,431 | -0.32152+ | 0.02147 |
| 3 | 24,970- | 1,599 | -0.32141+ | 0.01342 |
| 4 | 25,036+ | 1,919 | -0.32218- | 0.01643 |
| 5 | 25,070+ | 1,969 | -0.32232- | 0.01870 |
| 6 | 25,082+ | 5,689 | -0.32110+ | 0.05681 |
| 7 | 25,062+ | 7,267 | -0.32056+ | 0.06407 |
| 8 | 25,028+ | 4,950 | -0.32028+ | 0.03985 |
| 9 | 25,228+ | 5,847 | -0.32254- | 0.04967 |
| 10 | 25,344+ | 5,882 | -0.32306- | 0.05550 |
| 11 | 25,328+ | 9,602 | -0.32025+ | 0.09335 |
| 12 | 25,308+ | 12,147 | -0.31943+ | 0.10614 |
| 13 | 25,243+ | 8,617 | -0.31902+ | 0.06791 |
| 14 | 25,583+ | 10,025 | -0.32270- | 0.08370 |
| 15 | 25,798+ | 9,851 | -0.32371- | 0.09134 |
| WLSBF | | | | |
| Data Set 1 | 25,003+ | 1,805 | -0.32184+ | 0.01614 |
| 2 | 25,011+ | 1,443 | -0.32183+ | 0.01344 |
| 3 | 24,979- | 1,830 | -0.32148+ | 0.01666 |
| 4 | 25,033+ | 1,638 | -0.32216- | 0.01425 |
| 5 | 25,073+ | 1,626 | -0.32241- | 0.01538 |
| 6 | 25,124+ | 5,456 | -0.32155+ | 0.04862 |
| 7 | 25,157+ | 4,367 | -0.32156+ | 0.04031 |
| 8 | 25,050+ | 5,474 | -0.32047+ | 0.04903 |
| 9 | 25,222+ | 4,921 | -0.32252- | 0.04232 |
| 10 | 25,348+ | 4,835 | -0.32336- | 0.04574 |
| 11 | 25,404+ | 9,224 | -0.32108+ | 0.08123 |
| 12 | 25,469+ | 7,420 | -0.32200- | 0.06715 |
| 13 | 25,273+ | 9,154 | -0.31928+ | 0.08024 |
| 14 | 25,577+ | 8,305 | -0.32275- | 0.06978 |
| 15 | 25,799+ | 8,055 | -0.32417- | 0.07551 |

In all cases the WLSBF technique had a smaller 1st-3rd quartile range. Therefore, in cases 3, 8, and 13 the

presence of outliers caused the total range to be larger for WLSBF, but when the influence of the outliers is removed WLSBF was more efficient.

When the data was combined into three data sets the expectation was that the values obtained would be more reliable than they were previously. Or in other words, more confidence can be placed in the predictions made about the data. This was accomplished and the results are presented in Tables 13, 14, and 15. Data set one is the five data sets which were generated using a standard deviation of 0.04 (data sets 1-5), data set two is the five data sets generated using a standard deviation of 0.12 (data sets 6-10), and data set three is the five data sets generated using a standard deviation of 0.20 (data sets 11-15). The data presented in Tables 13, 14 and 15 are all averages. Recall that the letter A denotes the first unit cost and b is an expression related to the rate of learning or the slope coefficient.

In each of the combined data sets OLSBF outperformed WLSBF in predicting closer to the true value of the first unit cost. Notice that for both fitting techniques and for each data set, the estimated first unit cost parameters were greater than the true first unit cost. WLSBF was a more efficient estimator of the first unit cost in each of the three data sets if efficiency is measured as the dispersion between the first and third quartiles.

Table 13. Data Set One ($\sigma$ = 0.04)

| | OLSBF | | WLSBF | |
|---|---|---|---|---|
| | A | b | A | b |
| Maximum | 26,012 | -0.31265 | 25,893 | -0.31441 |
| Mean: | | | | |
| GM | 25,007 | -0.32183 | 25,018 | |
| AM | 25,009 | -0.32183 | 25,020 | -0.32194 |
| Minimum | 24,477 | -0.33050 | 24,225 | -0.32960 |
| Range: | | | | |
| Total | 1,535 | 0.01785 | 1,668 | 0.01519 |
| 1st-3rd Quartile | 505 | 0.00450 | 463 | 0.00424 |
| Bias: | | | | |
| GM Mean | 7 | | 18 | |
| AM Mean | 9 | 0.00010 | 20 | 0.00001 |

Table 14. Data Set Two ($\sigma$ = 0.12)

| | OLSBF | | WLSBF | |
|---|---|---|---|---|
| | A | b | A | b |
| Maximum | 28,308 | -0.29429 | 27,882 | -0.29955 |
| Mean: | | | | |
| GM | 25,124 | | 25,162 | |
| AM | 25,149 | -0.32151 | 25,180 | -0.32189 |
| Minimum | 22,381 | -0.34747 | 22,871 | -0.34476 |
| Range: | | | | |
| Total | 5,927 | 0.05318 | 5,011 | 0.04521 |
| 1st-3rd Quartiles | 1,498 | 0.01364 | 1,426 | 0.01258 |
| Bias: | | | | |
| GM Mean | 124 | | 162 | |
| AM Mean | 149 | 0.00042 | 180 | 0.00004 |

Table 15. Data Set Three ($\sigma$ = 0.20)

| | OLSBF | | WLSBF | |
|---|---|---|---|---|
| | A | b | A | b |
| Maximum | 31,029 | -0.27617 | 30,181 | -0.28494 |
| Mean: | | | | |
| GM | 25,381 | | 25,451 | |
| AM | 25,452 | -0.32102 | 25,504 | -0.32186 |
| Minimum | 20,980 | -0.36466 | 21,749 | -0.35972 |
| Range: | | | | |
| Total | 10,049 | 0.08849 | 8,432 | 0.07478 |
| 1st-3rd Quartiles | 2,511 | 0.02277 | 2,415 | 0.02098 |
| Bias: | | | | |
| GM Mean | 381 | | 451 | |
| AM Mean | 452 | 0.00091 | 504 | 0.00007 |

In predicting the slope coefficient parameter, WLSBF was not only the more efficient predictor, it also predicted closer to the true slope coefficient. In the worst case (data set three), the estimated slope coefficient parameter was only one-tenth of one percent from the true slope coefficient for the WLSBF technique. The bias in the estimated slope coefficient for the OLSBF was greater than or equal to ten times that of the bias in the estimated slope coefficient for WLSBF.

This leads to the next objective where the bias created by using the logarithmic transformation is reduced. The results from this objective indicate that bias is not introduced in the slope coefficient which agrees with the statement made by Daneman (10:6). On the other hand, bias is apparent in the first unit cost. This could be tested using the Wilcoxon sign ranked W-statistic (21:492-498), or by simply relating the occurance of overestimating the true value (+) as the head of a coin and underestimating the true value (-) as the tail on a coin. Data sets 1-5, 6-10, and 11-15 must be considered separatly since using the same seed value eliminated the independence of the data sets (data set 1, 6, and 11 are not independent, etc). The probability of having five overestimates in five observation would then be 0.05**5 or approximately three percent, given the data (or coin) is unbiased. This is the same value the Wilcoxon W-statistic returns. This unlikely event occurred for data sets 6-10 and 11-15 for both ordinary and weighted

least-squares which leads to rejecting the hypothesis that the data is unbiased at the 97 percent confidence level.

Research Objective 4

> Determine how well the estimated value for the intercept and slope approximate the true value. Determine whether the bias reduction factor proposed by Ilderton or the one proposed by Daneman is the most appropriate for learning curve data. Upon determining which bias reduction factor is most appropriate, determine if the factor's use will result in unbiased estimates.

The results obtained from the previous research objective strongly indicate that the direction of the bias in the estimated first unit cost parameter is on the high side, with only four of the thirty data sets predicting first unit costs below the true value. Of these four data sets all were generated using a standard deviation of 0.04. The same results strongly indicate that no bias exists in the estimate of the slope coefficient, where seventeen data sets have predicted values above the true value and thirteen data sets have predicted values below the true value. Thus, no correction factor was investigated for the predicted slope coefficient.

The bias reduction factor, for the estimated first unit cost, recommended by Ilderton (16:43-46) is in the correct direction in that it reduces the estimated value while the bias reduction factor recommended by Daneman for nonweighted regression would increase the bias that already exists. Only an approximation to the Ilderton bias reduction factor was used in this objective. The factor used was $\exp(-MSE/2)$ , where MSE is a measure of the variance of the

data, or the standard deviation squared. Therefore, the estimated first unit cost parameter in data sets 1-5 was multiplied by exp(-(0.04**2)/2) or 0.9992; the estimated first unit parameter in data sets 6-10 was multiplied by exp(-(0.12**2)/2) or 0.9928; and, the estimated first unit cost parameter in data sets 11-15 was multiplied by exp(-(0.20**2)/2) or 0.9802. Table 16 shows the results when the bias reduction factor is used. The same plus and minus sign notation as explained in the previous objective was used.

For OLSBF the reduced bias results predicted closer to the true first unit cost in seven of the fifteen data sets. For WLSBF the reduced bias results did much better, predictinng closer to the true first unit cost in twelve of the fifteen data sets. Comparing the OLSBF and WLSBF techniques, the later predicted closer to the true first unit cost in twelve of the fifteen reduced bias results.

The results from Table 12 showed only four data sets with estimated first unit costs below the true value. For OLSBF with the bias adjustment factor incorporated, there are now eighteen data sets with predicted first unit cost parameters below the true value. Using the coin flipping example again it is clear that three underestimates and two overestimates would lead towards acceptance that the data is not biased (remember data sets with like seed numbers are not independent). This was the case for WLSBF data sets 1-5, 6-10, and 11-15. The bias adjustment factor is apparently inappropriate for OLSBF since the results

indicate that the bias has simply been reversed from overestimating to underestimating the true first unit cost.

Table 16. Estimated First Unit Cost Parameters with Reduced Bias

| Fitting Technique / | First Unit Cost Parameter Estimate | |
|---|---|---|
| | Biased | Reduced Bias |
| OLSBF: | | |
| Data Set 1 | 24,990- | 24,970- |
| 2 | 24,981- | 24,961- |
| 3 | 24,970- | 24,950- |
| 4 | 25,036+ | 25,016+ |
| 5 | 25,070+ | 25,050+ |
| 6 | 25,082+ | 24,902- |
| 7 | 25,062+ | 24,882- |
| 8 | 25,028+ | 24,848- |
| 9 | 25,228+ | 25,047+ |
| 10 | 25,344+ | 25,162+ |
| 11 | 25,328+ | 24,826- |
| 12 | 25,308+ | 24,807- |
| 13 | 25,243+ | 24,743- |
| 14 | 25,583+ | 25,076+ |
| 15 | 25,798+ | 25,287+ |
| WLSBF: | | |
| Data Set 1 | 25,003+ | 24,983- |
| 2 | 25,011+ | 24,991- |
| 3 | 24,979- | 24,959- |
| 4 | 25,033+ | 25,013+ |
| 5 | 25,073+ | 25,053+ |
| 6 | 25,124+ | 24,944- |
| 7 | 25,157+ | 24,977- |
| 8 | 25,050+ | 24,870- |
| 9 | 25,222+ | 25,041+ |
| 10 | 25,348+ | 25,166+ |
| 11 | 25,404+ | 24,901- |
| 12 | 25,469+ | 24,965- |
| 13 | 25,273+ | 24,773- |
| 14 | 25,557+ | 25,071+ |
| 15 | 25,799+ | 25,288+ |

If any questions as to whether the data remained biased still exist Table 17 should remove all doubt. Here the data sets were once again combined. The reduced bias estimates for OLSBF are now all below the true value for the first

unit cost. The larger the standard deviation used in simulating the data set the farther below the true value. However, the reduced bias estimates for WLSBF are identical to the true first unit cost value regardless of the standard deviation used in simulating the data sets.

Table 17. A Comparison, on Average, of Biased and Reduced Biased Estimates for the First Unit Cost Parameters

| Fitting Technique | Standard Deviation | Biased Estimate | Reduced Bias Estimate |
|---|---|---|---|
| OLSBF | 0.04 | 25,009 | 24,989 |
| OLSBF | 0.12 | 25,149 | 24,968 |
| OLSBF | 0.20 | 25,452 | 24,948 |
| WLSBF | 0.04 | 25,020 | 25,000 |
| WLSBF | 0.12 | 25,180 | 25,000 |
| WLSBF | 0.20 | 25,500 | 25,000 |

Research Objective 5

Test whether the average unit cost data is normally distributed, in both the transformed and untransformed state.

The following null (Ho) and alternate (Ha) hypotheses were formulated in order to test whether the average unit cost data, $\ln(\bar{Y})$, are normally distributed. The level of significance for the test was chosen to be 0.10 and the test statistic was the Kolomogorov-Smirnov D-statistic.

Ho: The $\ln(\bar{Y})$'s are normally distributed
Ha: The $\ln(\bar{Y})$'s are not normally distributed

Ho will be accepted if the calculated D-statistic is less than or equal to 0.122, and will be rejected if the D-statistic is greater than 0.122. Rejection of the Ho leads to the conclusion that the statistics obtained from

use of the OLSBF and WLSBF techniques are flawed since the assumption of normality was violated. Table 18 shows the calculated D-statistics for the $\ln(\bar{Y})$'s for lot sizes 1, 8, 64, 512. The D-statistic for the $\bar{Y}$'s was also calculated. It was believed that the $\bar{Y}$'s would exhibit a nonnormal distribution if the $\ln(\bar{Y})$'s exhibited a normal distribution, and visa-versa. Note in Table 18, that this was not the case. In fact the calculated D-statistics seem to indicate that the distribution of the $\ln(\bar{Y})$'s and the $\bar{Y}$'s are similar. The D-statistics for both remain below the critical value, and when the calculated D-statistic increases or decreases for one the same trend occurs in the other. These results would lead to acceptance of Ho, meaning that the $\ln(\bar{Y})$'s and the $\bar{Y}$'s are normally distributed. This peculiarity beckons further research.

Table 18. D-Statistic for $\ln(\bar{Y})$'s and $\bar{Y}$'s

| Lot Size | D-$\ln(\bar{Y})$'s | D-$\bar{Y}$'s |
|---|---|---|
| 1 | 0.056855 | 0.072321 |
| 8 | 0.048555 | 0.051683 |
| 64 | 0.100895 | 0.101462 |
| 512 | 0.080249 | 0.079414 |

Research Objective 6

> Test to see whether the weighting scheme used in the ICLOT program and the Hutchison learning curve program can be improved upon.

The objective here was to calculate the variance of production lot data for lots which incrementally increased in size. Recall that the variance for a lot of size n is

$\sigma^2/n$. Therefore, the hypothsis was that if the weighting scheme used in ICLOT and the Hutchison learning curve program was correct then the variance for a lot of size n would be twice the variance of a lot of size 2*n. The calculated variance of the $\ln(\bar{Y})$'s and the $\bar{Y}$'s were for lots of size 1, 8, 64, and 512. There were 100 $\ln(\bar{Y})$'s and $\bar{Y}$'s simulated for each lot size. Table 19 shows the calculated variance for each data set.

Table 19. Calculated Variances for $\ln(\bar{Y})$'s and Y's Using a Multiplicative Error Term

| Lot Size | Variance-$\ln(\bar{Y})$'s | Variance-$\bar{Y}$'s |
|---|---|---|
| 1 | 0.0160000 | 10715181.79 |
| 8 | 0.0019500 | 359437.84 |
| 64 | 0.0002294 | 11328.92 |
| 512 | 0.0000283 | 367.08 |

If the data is homoscedastic the variance will decrease by the inverse of the increase in the lot size. For example, the variance for the data from lot sizes of eight would be one-eight that of the variance of the data from lot sizes of one given that the data is homoscedastic. When the data was generated using a multiplicative error term the variance of the $\ln(\bar{Y})$'s decreases by a factor of 8.2, 8.5, and 8.1 for lot sizes 1 to 512 respectively. For the $\bar{Y}$'s the variance decreases by a factor of 29.8, 31.7, and 30.9 for lot sizes 1 to 512 respectively. By generating the data with an additive error term the results dramatically change as Table 20 depicts. Here the variance of the $\ln(\bar{Y})$'s decreases by a factor of 3.1, 2.3, and 2.2 for lot sizes 1 to 512

respectively, and the variance of the $\bar{Y}$'s decreases by a factor of 10.8, 8.4, and 8.3 for lot sizes 1 to 512 respectively.

Table 20. Calculated Variance for $\ln(\bar{Y})$'s and $\bar{Y}$'s Using an Additive Error Term

| Lot Size | Variance-$\ln(\bar{Y})$'s | Variance-$\bar{Y}$'s |
|---|---|---|
| 1 | 0.00044300 | 278382.44 |
| 8 | 0.00014290 | 25855.33 |
| 64 | 0.00006304 | 3069.88 |
| 512 | 0.00002893 | 370.29 |

If weighting by lot size is correct for the ln(Y)'s generated using the multiplicative error term, the expected variance for lots of size 1, 8, 64, and 512 would be 0.0144/n respectively. The lot data simulated had variances relatively close to the expected variance as Table 19 shows. This did not hold true for the $\ln(\bar{Y})$'s generated using the additive error term. Here the expected variance was 0.004/n . Only the $\ln(\bar{Y})$ for lot size one was close, with the variance of the remaining $\ln(\bar{Y})$'s exhibiting reductions by factors between two and three instead of eight.

Note that the $\bar{Y}$'s generated with the additive error term show a reduction of approximately a factor of eight. However, when lot data is generated with the multiplicative error term, the variance is reduced by factors of between 29 and 31. The expected variance of the $\bar{Y}$'s generated with the multiplicative error term is 10,159,654/n , which was approximated only when the lot size was one. The expected

variance of the $\bar{Y}$'s generated with the additive error term is $250,000/n$, which as mentioned above was closely approximated.

Chapter VI which follows drew conclusions from the six Research Objectives where possible. Where no conclusion could be drawn, recommendations for further research were made.

# VI. Conclusions and Recommendations

## Chapter Overview

This chapter covers the conclusions of this research based on the results of the six research objectives. The conclusions will then be followed with some recommendations for further research pertaining to the topic of learning curves.

## Summary of Findings

The initial part of this research involved a comparison of the true lot midpoint algorithm to the lot plot point heuristic. Use of the lot plot point heuristic created bias in both the estimated first unit cost parameter for both ordinary and weighted least-squares in excess of the true lot midpoint algorithm. Yet, bias still exists. Use of the lot plot point heuristic, unlike true lot midpoint, created bias in the slope coefficient when using the WLSBF technique. No bias exists in the slope coefficient for either ordinary or weighted least-squares when the true lot midpoint algorithm was used. Therefore, the slope coefficient parameter estimated by the ICLOT program and the Hutchison learning curve program is correct while the estimated slope coefficient for both programs is biased on the high side.

An approximation of the bias reduction factor recommended by Ilderton was tested. Ilderton's adjustment factor was based mainly on the amount of variance contained

in the fitted data while the approximate bias adjustment factor was based entirely on the variance of the data. The formula used to reduce the bias in the estimate of the first unit cost parameter was exp(-MSE/2) . Average biases of two percent were present in the estimated first unit cost parameters when the standard deviation used in simulating the data was 0.20. The biases were significantly less when the standard deviations were 0.12 and 0.04. For most DoD programs this bias would result in small overestimates of cost. These overestimates may be tolerated, however, they are unnecessary if the analyst is aware of the presence of bias in the first unit cost parameter.

Following is a table which presents scenarios of the effects that this bias may have on DoD programs. Understand that the estimated cost of the first unit affects the cost estimate in many areas of a program besides hardware costs. It affects the remaining hardware costs and any costs which are factored off that hardware. In the example the unbiased first unit cost is ten million dollars, the learning curve slope is 90 percent, and the number of production units is 1,000.

Table 21. Effect of Bias on an Estimate

| | Hardware Estimate | | |
|---|---|---|---|
| MSE | Biased | Unbiased | Overestimate |
| 0.05**2 | $4,126,873,731 | $4,121,718,360 | $ 5,155,371 |
| 0.10**2 | 4,142,378,560 | 4,121,718,360 | 20,660,200 |
| 0.15**2 | 4,168,349,499 | 4,121,718,360 | 46,631,139 |
| 0.20**2 | 4,204,982,594 | 4,121,718,360 | 83,264,234 |

With the budget cuts currently being experienced by the DoD, programs can not afford unnecessary overestimates.

The techniques of OLSBF and WLSBF were compared with a surprising result. Due to the bias in the estimated first unit cost parameter OLSBF predicted closer to the true value than did WLSBF. In general WLSBF was a more efficient estimator, except that (for the data simulated) OLSBF had a smaller prediction range than WLSBF when a standard deviation of 0.04 was used in the data simulation. So for analysts using the Hutchison learning curve program, with no intent to use the bias reduction factor, the best point estimate is obtained by using OLSBF while a tighter confidence band about the point estimate would be obtained using WLSBF.

If the bias reduction factor is used the WLSBF outperforms OLSBF for both efficiency and accuracy of prediction. The recommendation is to use the WLSBF technique (automatic with ICLOT) to get the first unit cost estimate, then adjust for the bias.

After running the tests for normality and constant variance it seems reasonable, to assume that the data is normally distributed and that weighting by lot size is the appropriate weighting scheme for a learning curve model based on a multiplicative error term. If the learning curve model is based on an additive error term both the linear model and the weighting scheme would require investigation.

Areas for Further Research

The most promising area for further research is in the learning curve model itself. Variables which result in cost reduction, besides cumulative quantity should be investigated. While cumulative quantity certainly provides workers and managers with the opportunity to "learn" are there variables in a production effort which more closely relate to the rate of cost reductions? Factors such as production engineering, production rate, quality assurance, technology, and management philosophy may be the key to not only predicting the rate of learning but in actively managing the rate of learning.

Developing programs which better fit a curve to production data is another wide open area for further study. This research showed that programming the bias reduction factor into the ICLOT and Hutchison learning curve programs (when the WLSBF option is chosen) would be an important step to providing better program cost estimates. Calculating a bias adjustment factor for the OLSBF technique is still needed. This research showed that the bias adjustment factor, exp(-MSE/2) , only worked for the first unit cost parameters obtained using the WLSBF technique. It merely reversed the bias in the estimated first unit cost parameters obtained using the OLSBF technique. Investigating the potential use for non-linear regression and nonparametric regression in fitting learning curve data could be fruitful, along with developing programs which utilize these fitting techniques. Developing programs which

allow the user to fit learning curves such as the Stanford-B model (Y = A * (X +B)**b) , a model which takes into account the contractor's prior experience on similar programs; and, models which take into account production breaks could prove useful in the cost estimating community.

Finally, further examination into actually learning curve data should be accomplished. This may answer questions about autocorrelation in learning curve data, which violates the assumption of independence. Adding the Durbin-Watson test statistic to learning curve programs would provide the informed analyst with information on potential independence problems which might warrant further examination of the data. Examination as whether the multiplicative error term (constant percentage variance) or the additive error term (constant variance) is most appropriate. If the multiplicative error term is more appropriate giving added weight to lots later in the production process makes sense, since both contractor and government have gained significant experience in costing each system.

# APPENDIX A

## Sample SAS Programs for Data Simulation

These programs were written and run using the VMS Version of SAS, Release 5.16. Reference SAS Institute Inc. (1985).

Program I - This program simulates unit cost learning curve data in logarithmic form, given a learning curve slope of 80 percent, a first unit cost of 25,000, and an error term of 0.12. The data is then tranformed to it's standard state using the antilogarithm fuction. One hundred production runs with 210 units per production run are generated. The output is unit number, cost per unit, and cumulative total cost.

```
 1        DATA ONE;
 2        A = LOG(25000);
 3        B = LOG(.8)/LOG(2);
 4        C = .12;
 5        DO J = 1 TO 100;
 6          TOTCOST = 0;
 7            DO I = 1 TO 210;
 8            LNX  = LOG(I);
 9            Z = RANNOR(1446);
10            LNY = A + (B * LNX) + (C * Z);
11            COST = EXP(LNY);
12            TOTCOST = TOTCOST + COST;
13            FILE NORMAL;
14            PUT I COST TOTCOST;
15            END;
16        END;
17        PROC PRINT;
```

Program II - This program simulates production lot data using the data from Program I of Appendix A. A 7x5 matrix is set up to divide each production run into seven lots, with the following lot data in each matrix: cumulative units, cumulative cost, lot cost, heuristic lot plot point, and average unit cost per lot. One hundred such matrices are generated.

```
 1        DATA TWO;
 2        INFILE NORMAL;
 3        INPUT N1-N630;
 4        ARRAY NUM[630] N1-N630;
 5        ARRAY COST[210] YY1-YY210;
 6        ARRAY LOT[7,5] L1-L35;
 7        ARRAY TOTCOST[210] TC1-TC210;
 8          DO I = 1 TO 210;
 9            COST[I] = NUM[I * 3 - 1];
10            TOTCOST[I] = NUM[I * 3];
11          END;
12        LABEL1: TEMP = 10 * RANUNI(1515);
13                TEMP = ROUND(TEMP,1);
14                IF TEMP> 10 OR TEMP< 2 THEN GO TO LABEL1;
15            LOT[1,1] = TEMP;
16            LOT[1,2] = TOTCOST[TEMP];
17            LOT[1,3] = TOTCOST[TEMP];
18            IF TEMP = 10 THEN LOT[1,4] = LOT[1,1]/3;
19                         ELSE LOT[1,4] = LOT[1,1]/2;
20            LOT[1,5] = LOT[1,3]/TEMP;
21        LABEL2: TEMP = 100 * RANUNI(1515);
22                TEMP = ROUND(TEMP,1);
23                IF TEMP> 25 OR TEMP< 15 THEN GO TO LABEL2;
24            LOT[2,1] = TEMP + LOT[1,1];
25            LOT[2,2] = TOTCOST[LOT[2,1]];
26            LOT[2,3] = LOT[2,2] - LOT[1,2];
27            LOT[2,4] = TEMP/2 + LOT[[1,1];
28            LOT[2,5] = LOT[2,3]/TEMP;
29        LABEL3: TEMP = 100 * RANUNI(1515);
30                TEMP = ROUND(TEMP,1);
31                IF TEMP> 30 OR TEMP< 20 THEN GO TO LABEL3;
32            LOT[3,1] = TEMP.+ LOT[2,1];
33            LOT[3,2] = TOTCOST[LOT[3,1]];
34            LOT[3,3] = LOT[3,2] - LOT[2,2];
35            LOT[3,4] = TEMP/2 + LOT[2,1];
36            LOT[3,5] = LOT[3,3]/TEMP;
37        LABEL4: TEMP = 100 * RANUNI(1515);
38                TEMP = ROUND(TEMP,1);
39                IF TEMP> 35 OR TEMP< 25 THEN GO TO LABEL4;
40            LOT[4,1] = TEMP + LOT[3,1];
41            LOT[4,2] = TOTCOST[LOT[4,1]];
42            LOT[4,3] = LOT[4,2] - LOT[3,2];
43            LOT[4,4] = TEMP/2 + LOT[3,1];
44            LOT[4,5] = LOT[4,3]/TEMP;
45        LABEL5: TEMP = 100 * RANUNI(1515);
46                TEMP = ROUND(TEMP,1);
```

```
47                IF TEMP> 40 OR TEMP< 30 THEN GO TO LABEL5;
48              LOT[5,1] = TEMP + LOT[4,1];
49              LOT[5,2] = TOTCOST[LOT[5,1]];
50              LOT[5,3] = LOT[5,2] - LOT[4,2];
51              LOT[5,4] = TEMP/2 + LOT[4,1];
52              LOT[5,5] = LOT[5,3]/TEMP;
53          LABEL6: TEMP = 100 * RANUNI(1515);
54                TEMP = ROUND(TEMP,1);
55                IF TEMP> 50 OR TEMP< 40 THEN GO TO LABEL6;
56              LOT[6,1] = TEMP + LOT[5,1];
57              LOT[6,2] = TOTCOST[LOT[6,1]];
58              LOT[6,3] = LOT[6,2] - LOT[5,2];
59              LOT[6,4] = TEMP/2 + LOT[5,1];
60              LOT[6,5] = LOT[6,3]/TEMP;
61          LABEL7: TEMP = 210 - LOT[6,1];
62              LOT[7,1] = TEMP + LOT[6,1];
63              LOT[7,2] = TOTCOST[LOT[7,1]];
64              LOT[7,3] = LOT[7,2] - LOT[6,2];
65              LOT[7,4] = TEMP/2 + LOT[6,1];
66              LOT[7,5] = LOT[7,3]/TEMP;
66          FILE ULOTGEN;
67            DO I = 1 TO 7;
68              PUT LOT[I,1] LOT[I,2] LOT[I,3] LOT[I,4]
                  LOT[I,5];
69            END;
```

Program III - This program simulates production lot data using the data from Program I of Appendix A. A 7x6 matrix is set up to divide the each production run into seven lots, with the following lot data in each matrix: cumulative units, cumulative cost, lot cost, true lot midpoint, average unit cost per lot, and lot size. The difference between this program and Program II is true lot midpoint, and the inclusion of a column for lot size. One hundred such matrices are generated.

```
 1      DATA TWO;
 2      INFILE NORMAL;
 3      INPUT N1-N630;
 4      ARRAY NUM[630] N1-N630;
 5      ARRAY COST[210] YY1-YY210;
 6      ARRAY LOT[7,6] L1-L42;
 7      ARRAY TOTCOST[210] TC1-TC210;
 8        DO I = 1 TO 210;
 9          COST[I] = NUM[I * 3 - 1];
10          TOTCOST[I] = NUM[I * 3];
11        END;
12      LABEL1: TEMP = 10 * RANUNI(1515);
13              TEMP = ROUND(TEMP,1);
14              IF TEMP> 10 OR TEMP< 2 GO TO LABEL1;
15          LOT[1,1] = TEMP;
16          LOT[1,2] = TOTCOST[TEMP];
17          LOT[1,3] = TOTCOST[TEMP];
18          LOT[1,4] = 0;
19          B = LOG(.8)/LOG(2);
20            DO I = 1 TO LOT[1,1];
21              DUM = LOT[1,4] + (I**B);
22              LOT[1,4] = DUM;
23            END;
24          LOT[1,4] = (LOT[1,4]/LOT[1,1])**(1/B);
25          LOT[1,5] = LOT[1,3]/TEMP;
26          LOT[1,6] = LOT[1,1];
27      LABEL2: TEMP = 100 * RANUNI(1515);
28              TEMP = ROUND(TEMP,1);
29              IF TEMP> 25 OR TEMP< 15 THEN GO TO LABEL2;
30          LOT[2,1] = TEMP + LOT[1,1];
31          LOT[2,2] = TOTCOST[LOT[2,1]];
32          LOT[2,3] = LOT[2,2] - LOT[1,2];
33          LOT[2,4] = 0;
34          Z = LOT[2,1] - LOT[1,1];
35            DO I = 1 TO Z;
36              DUM = LOT[2,4] + ((I + LOT[1,1])**B);
37              LOT[2,4] = DUM;
38            END;
39          LOT[2,4] = (LOT[2,4]/Z)**(1/B);
40          LOT[2,5] = LOT[2,3]/TEMP;
41          LOT[2,6] = LOT[2,1] - LOT[1,1];
42      LABEL3: TEMP = 100 * RANUNI(1515);
43              TEMP = ROUND(TEMP,1);
44              IF TEMP> 30 OR TEMP< 20 THEN GO TO LABEL3;
```

```
45          LOT[3,1] = TEMP + LOT[2,1];
46          LOT[3,2] = TOTCOST[LOT[3,1]];
47          LOT[3,3] = LOT[3,2] - LOT[2,2];
48          LOT[3,4] = 0;
49          Z = LOT[3,1] - LOT[2,1];
50            DO I = 1 TO Z;
51              DUM = LOT[3,4] + ((I + LOT[2,1])**B);
52              LOT[3,4] = DUM;
53            END;
54          LOT[3,4] = (LOT[3,4]/Z)**(1/B);
55          LOT[3,5] = LOT[3,3]/TEMP;
56          LOT[3,6] = LOT[3,1] - LOT[2,1];
57      LABEL4: TEMP = 100 * RANUNI(1515);
58              TEMP = ROUND(TEMP,1);
59              IF TEMP> 35 OR TEMP< 25 THEN GO TO LABEL4;
60          LOT[4,1] = TEMP + LOT[3,1];
61          LOT[4,2] = TOTCOST[LOT[4,1]];
62          LOT[4,3] = LOT[4,2] - LOT[3,2];
63          LOT[4,4] = 0;
64          Z = LOT[4,1] - LOT[3,1];
65            DO I = 1 TO Z;
66              DUM = LOT[4,4] + ((I + LOT[3,1])**B);
67              LOT[4,4] = DUM;
68            END;
69          LOT[4,4] = (LOT[4,4]/Z)**(1/B);
70          LOT[4,5] = LOT[4,3]/TEMP;
71          LOT[4,6] = LOT[4,1] - LOT[3,1];
72      LABEL5: TEMP = 100 * RANUNI(1515);
73              TEMP = ROUND(TEMP,1);
74              IF TEMP> 40 OR TEMP< 30 THEN GO TO LABEL5;
75          LOT[5,1] = TEMP + LOT[4,1];
76          LOT[5,2] = TOTCOST[LOT[5,1]];
77          LOT[5,3] = LOT[5,2] - LOT[4,2];
78          LOT[5,4] = 0;
79          Z = LOT[5,1] - LOT[4,1];
80            DO I = 1 TO Z;
81              DUM = LOT[5,4] + ((I + LOT[4,1])**B);
82              LOT[5,4] = DUM;
83            END;
84          LOT[5,4] = (LOT[5,4]/Z)**(1/B);
85          LOT[5,5] = LOT[5,3]/TEMP;
86          LOT[5,6] = LOT[5,1] - LOT[4,1];
87      LABEL6: TEMP = 100 * RANUNI(1515);
88              TEMP = ROUND(TEMP,1);
89              IF TEMP> 50 OR TEMP< 40 THEN GO TO LABEL6:
90          LOT[6,1] = TEMP + LOT[5,1];
91          LOT[6,2] = TOTCOST[LOT[6,1]];
92          LOT[6,3] = LOT[6,2] - LOT[5,2];
93          LOT[6,4] = 0;
94          Z = LOT[6,1] - LOT[5,1];
95            DO I = 1 TO Z;
96              DUM = LOT[6,4] + ((I + LOT[5,1])**B);
97              LOT[6,4] = DUM;
98            END;
99          LOT[6,4] = (LOT[6,4]/Z)**(1/B);
```

```
100           LOT[6,5] = LOT[6,3]/TEMP;
101           LOT[6,6] = LOT[6,1] - LOT[5,1];
102       TEMP = 210 - LOT[6,1];
103           LOT[7,1] = 210;
104           LOT[7,2] = TOTCOST[LOT[7,1]];
105           LOT[7,3] = LOT[7,2] - LOT[6,2];
106           LOT[7,4] = 0;
107           Z = LOT[7,1] - LOT[6,1];
108             DO I = 1 TO Z;
109               DUM = LOT[7,4] + ((I + LOT[6,1])**B);
110               LOT[7,4] = DUM;
111             END;
112           LOT[7,4] = (LOT[7,4]/Z)**(1/B);
113           LOT[7,5] = LOT[7,3]/TEMP;
114           LOT[7,6] = LOT[7,1] - LOT[6,1];
115         FILE ULOTGEN2;
116           DO I = 1 TO 7;
117             PUT LOT[I,1] LOT[I,2] LOT[I,3] LOT[I,4]
                      LOT[I,5] LOT[I,6];
118           END;
```

Program IV - This program simulates production lot data using the data from Program I of Appendix A. A 7x6 matrix is set up to divide each production run into seven lots, with the following data in each matrix: cumulative units, cumulative cost, lot cost, true lot midpoint, average unit cost per lot, and lot size. The difference between this program and Program III is the lot size generation. One hundred such matrices are generated.

```
 1        DATA TWO;
 2        INFILE NORMAL;
 3        INPUT N1-N630;
 4        ARRAY NUM[630] N1-N630;
 5        ARRAY COST[210] YY1-YY210;
 6        ARRAY LOT[7,6] L1-L42;
 7        ARRAY TOTCOST[210] TC1-TC210;
 8          DO I = 1 TO 210;
 9            COST[I] = NUM[I * 3 - 1];
10            TOTCOST[I] = NUM[I * 3];
11          END;
12        TEMP = 5 + (5 * RANUNI(1515));
13        TEMP = ROUND(TEMP,1);
14          LOT[1,1] = TEMP;
15          LOT[1,2] = TOTCOST[TEMP];
16          LOT[1,3] = TOTCOST[TEMP];
17          LOT[1,4] = 0;
18          B = LOG(.8)/LOG(2);
19            DO I = 1 TO LOT[1,1];
20              DUM = LOT[1,4] + (I**B);
21              LOT[1,4] = DUM;
22            END;
23          LOT[1,4] = (LOT[1,4]/LOT[1,1])**(1/B);
24          LOT[1,5] = LOT[1,3]/TEMP;
25          LOT[1,6] = LOT[1,1];
26        TEMP = 15 + (5 * RANUNI(1515));
27        TEMP = ROUND(TEMP,1);
28          LOT[2,1] = TEMP + LOT[1,1];
29          LOT[2,2] = TOTCOST[LOT[2,1]];
30          LOT[2,3] = LOT[2,2] - LOT[1,2];
31          LOT[2,4] = 0;
32          Z = LOT[2,1] - LOT[1,1];
33            DO I = 1 TO Z;
34              DUM = LOT[2,4] + ((I + LOT[1,1])**B);
35              LOT[2,4] = DUM;
36            END;
37          LOT[2,4] = (LOT[2,4]/Z)**(1/B);
38          LOT[2,5] = LOT[2,3]/TEMP;
39          LOT[2,6] = LOT[2,1] - LOT[1,1];
40        TEMP = 25 + (10 * RANUNI(1515));
41        TEMP = ROUND(TEMP,1);
42          LOT[3,1] = TEMP + LOT[2,1];
43          LOT[3,2] = TOTCOST[LOT[3,1]];
44          LOT[3,3] = LOT[3,2] - LOT[2,2];
45          LOT[3,4] = 0;
```

```
46          Z = LOT[3,1] - LOT[2,1];
47            DO I = 1 TO Z;
48              DUM = LOT[3,4] + ((I + LOT[2,1])**B);
49              LOT[3,4] = DUM;
50            END;
51          LOT[3,4] = (LOT[3,4]/Z)**(1/B);
52          LOT[3,5] = LOT[3,3]/TEMP;
53          LOT[3,6] = LOT[3,1] - LOT[2,1];
54        TEMP = 40 + (10 * RANUNI(1515));
55        TEMP = ROUND(TEMP,1);
56          LOT[4,1] = TEMP + LOT[3,1];
57          LOT[4,2] = TOTCOST[LOT[4,1]];
58          LOT[4,3] = LOT[4,2] - LOT[3,2];
59          LOT[4,4] = 0;
60          Z = LOT[4,1] - LOT[3,1];
61            DO I = 1 TO Z;
62              DUM = LOT[4,4] + ((I + LOT[3,1])**B);
63              LOT[4,4] = DUM;
64            END;
65          LOT[4,4] = (LOT[4,4]/Z)**(1/B);
66          LOT[4,5] = LOT[4,3]/TEMP;
67          LOT[4,6] = LOT[4,1] - LOT[3,1];
68        TEMP = 40 + (10 * RANUNI(1515));
69        TEMP = ROUND(TEMP,1);
70          LOT[5,1] = TEMP + LOT[4,1];
71          LOT[5,2] = TOTCOST[LOT[5,1]];
72          LOT[5,3] = LOT[5,2] - LOT[4,2];
73          LOT[5,4] = 0;
74          Z = LOT[5,1] - LOT[4,1];
75            DO I = 1 TO Z;
76              DUM = LOT[5,4] + ((I + LOT[4,1])**B);
77              LOT[5,4] = DUM;
78            END;
79          LOT[5,4] = (LOT[5,4]/Z)**(1/B);
80          LOT[5,5] = LOT[5,3]/TEMP;
81          LOT[5,6] = LOT[5,1] - LOT[4,1];
82        TEMP = 40 + (10 * RANUNI(1515));
83        TEMP = ROUND(TEMP,1);
84          LOT[6,1] = TEMP + LOT[5,1];
85          LOT[6,2] = TOTCOST[LOT[6,1]];
86          LOT[6,3] = LOT[6,2] - LOT[5,2];
87          LOT[6,4] = 0;
88          Z = LOT[6,1] - LOT[5,1];
89            DO I = 1 TO Z;
90              DUM = LOT[6,4] + ((I + LOT[5,1])**B);
91              LOT[6,4] = DUM;
92            END;
93          LOT[6,4] = (LOT[6,4]/Z)**(1/B);
94          LOT[6,5] = LOT[6,3]/TEMP;
95          LOT[6,6] = LOT[6,1] - LOT[5,1];
96        TEMP = 210 - LOT[6,1];
97          LOT[7,1] = 210;
98          LOT[7,2] = TOTCOST[LOT[7,1]];
99          LOT[7,3] = LOT[7,2] - LOT[6,2];
100         LOT[7,4] = 0;
```

```
101           Z = LOT[7,1] - LOT[6,1];
102             DO I = 1 TO Z;
103               DUM = LOT[7,4] + ((I + LOT[6,1])**B);
104               LOT[7,4] = DUM;
105             END;
106           LOT[7,4] = (LOT[7,4]/Z)**(1/B);
107           LOT[7,5] = LOT[7,3]/TEMP;
108           LOT[7,6] = LOT[7,1] - LOT[6,1];
109         FILE ULOTGEN2;
110           DO I = 1 TO 7;
111             PUT LOT[I,1] LOT[I,2] LOT[I,3] LOT[I,4]
                    LOT[I,5] LOT[I,6];
112           END;
```

Program V - This program simulates unit cost learning curve data in logarithmic form, given a learning curve slope of 80 percent, a first unit cost of 25,000, and an error term of 0.12. The data is then transformed to it's standard state using the antilogarithm function. One hundred production runs of 585 units per production run are generated. The output is unit number, cost per unit, and cumulative total cost.

```
 1        DATA ONE;
 2        A = LOG(25000);
 3        B = LOG(.8)/LOG(2);
 4        C = .12;
 5        DO J = 1 TO 100;
 6          TOTCOST = 0;
 7            DO I = 1 TO 585;
 8            LNX  = LOG(I);
 9            Z = RANNOR(1111);
10            LNY = A + (B * LNX) + (C * Z);
11            COST = EXP(LNY);
12            TOTCOST = TOTCOST + COST;
13            FILE NORMAL1;
14            PUT I COST TOTCOST;
15            END;
16        END;
17        PROC PRINT;
```

Program VI - This program simulates production lot data using the data from Program V or VII of Appendix A. A 4x6 matrix is set up to divide each production run into four lots, with the following lot data in each matrix: cumulative units, cumulative cost, lot cost, true lot midpoint, average unit cost per lot, and lot size. One hundred such matrices are generated.

```
 1        DATA TWO;
 2        INFILE NORMAL1;
 3        INPUT N1-N1755;
 4        ARRAY NUM[1755] N1-N1755;
 5        ARRAY COST[210] YY1-YY585;
 6        ARRAY LOT[4,6] L1-L24;
 7        ARRAY TOTCOST[585] TC1-TC585;
 8          DO I = 1 TO 585;
 9            COST[I] = NUM[I * 3 - 1];
10            TOTCOST[I] = NUM[I * 3];
11          END;
12          LOT[1,1] = 1;
13          LOT[1,2] = TOTCOST[1];
14          LOT[1,3] = TOTCOST[1],
15          LOT[1,4] = 0;
16          B = LOG(.8)/LOG(2);
17            DO I = 1 TO LOT[1,1];
18              DUM = LOT[1,4] + (I**B);
19              LOT[1,4] = DUM;
20            END;
21          LOT[1,4] = (LOT[1,4]/LOT[1,1])**(1/B);
22          LOT[1,5] = LOT[1,3]/TEMP;
23          LOT[1,6] = LOT[1,1];
24          LOT[2,1] = 9;
25          LOT[2,2] = TOTCOST[LOT[2,1]];
26          LOT[2,3] = LOT[2,2] - LOT[1,2];
27          LOT[2,4] = 0;
28          Z = LOT[2,1] - LOT[1,1];
29            DO I = 1 TO Z;
30              DUM = LOT[2,4] + ((I + LOT[1,1])**B);
31              LOT[2,4] = DUM;
32            END;
33          LOT[2,4] = (LOT[2,4]/Z)**(1/B);
34          LOT[2,5] = LOT[2,3]/9;
35          LOT[2,6] = LOT[2,1] - LOT[1,1];
36          LOT[3,1] = 73;
37          LOT[3,2] = TOTCOST[LOT[3,1]];
38          LOT[3,3] = LOT[3,2] - LOT[2,2];
39          LOT[3,4] = 0;
40          Z = LOT[3,1] - LOT[2,1];
41            DO I = 1 TO Z;
42              DUM = LOT[3,4] + ((I + LOT[2,1])**B);
43              LOT[3,4] = DUM;
44            END;
45          LOT[3,4] = (LOT[3,4]/Z)**(1/B);
46          LOT[3,5] = LOT[3,3]/73;
```

```
47          LOT[3,6] = LOT[3,1] - LOT[2,1];
48          LOT[4,1] = 585;
49          LOT[4,2] = TOTCOST[LOT[4,1]];
50          LOT[4,3] = LOT[4,2] - LOT[3,2];
51          LOT[4,4] = 0;
52          Z = LOT[4,1] - LOT[3,1];
53            DO I = 1 TO Z;
54              DUM = LOT[4,4] + ((I + LOT[3,1])**B);
55              LOT[4,4] = DUM;
56            END;
57          LOT[4,4] = (LOT[4,4]/Z)**(1/B);
58          LOT[4,5] = LOT[4,3]/585;
59          LOT[4,6] = LOT[4,1] - LOT[3,1];
60        FILE ULOTGEN3;
61          DO I = 1 TO 4;
62            PUT LOT[I,1] LOT[I,2] LOT[I,3] LOT[i,4]
                                      LOT[I,5] LOT[I,6];
63          END;
```

Program VII - This program simulates unit cost learning curve data in standard form, given a learning curve slope of 80 percent, a first unit cost of 25,000, and an additive error term equal to 500 or two percent of the first unit cost. One hundred production runs of 585 units per production run are generated. The output is unit number, cost per unit, and cumulative total cost.

```
 1          DATA ONE;
 2          A = 25000;
 3          B = LOG(.8)/LOG(2);
 4          C = 500;
 5          DO J = 1 TO 100;
 6            TOTCOST = 0;
 7              DO I = 1 TO 585;
 8              X  = I;
 9              Z = RANNOR(1111);
10              Y = A * (X** B) + (C * Z);
11              COST = Y;
12              TOTCOST = TOTCOST + COST;
13              FILE NORMAL;
14              PUT I COST TOTCOST;
15              END;
16         END;
17         PROC PRINT;
```

# APPENDIX B

## Sample SAS Programs for Data Analysis

These programs were written and run using the VMS Version of SAS, Release 5.16. Reference SAS Institute Inc. (1985).

Program I - This program performs OLSBF on the data generated in Programs II, III, and IV of Appendix A (on line 4 the column LOTSZ does not exist in Program II of Appendix A, thus it would be deleted from the program). The data used is heuristic lot plot point or true lot midpoint (LPP), and the average unit cost per lot (AVUNCST). This data is transformed using the natural logarithm function (LOG) and then the technique of OLS fits the data. A permanent SAS data file is created which contains the intercept and slope coefficient data which will be used later. A list file is also generated which contains ANOVA tables for each production run (100).

```
 1        LIBNAME MINE '[TTRACHT]';
 2        DATA THREE;
 3        INFILE ULOTGEN;
 4          INPUT CUMX TOTY LOTCOST LPP AVUNCST LOTSZ;
 5        INFILE PRODRUN;
 6          INPUT PRDRUN;
 7        LNLPP = LOG(LPP);
 8        LNAVUCST = LOG(AVUNCST);
 9        PROC REG OUTEST = MINE.PARAMS;
10          MODEL LNAVUCST = LNLPP;
11            BY PRDRUN;
12        PROC PRINT;
```

Program II - This program performs WLSBF on the data generated in Programs II and III of Appendix A. This program was written by Avinger. Here the X, which is the weights, is equal to CUMX. The process and the output is the same as above program.

```
 1        LIBNAME MINE '[TTRACHT]';
 2        DATA THREE;
 3        HOLDX = 210;
 4        INFILE ULOTGEN;
 5          INPUT CUMX TOTY LOTCOST LPP AVUNCST LOTSZ;
 6        INFILE PRODRUN;
 7          INPUT PRDRUN;
 8        LNLPP = LOG(LPP);
 9        LNAVUCST = LOG(AVUNCST);
10          IF HOLDX = 210 THEN X = CUMX;
11                         ELSE X = CUMX - HOLDX;
12        PROC REG OUTEST = MINE.WPARAMS;
13          MODEL LNAVUCST = LNLPP;
14            WEIGHT X;
15            BY PRDRUN;
16        PROC PRINT;
```

Program III - This program take the slope coefficient (LNLPP) and first unit cost parameter (INTERCEP) from Programs I, II, and V, transforms them back to their standard state using the antilogarithm function. By tranforming each parameter the arithmetic mean of these parameters can be calculated, which is what PROC MEANS MEAN accomplishes for the variables SLOPE and EXPINTER. Also calculated is the geometric mean of the INTERCEP and LNLPP variables.

```
1         LIBNAME SDAT '[TTRACHT]';
2         DATA FOUR;
3         SET SDAT.PARAMS;
4         EXPINTER = EXP(INTERCEP);
5         SLOPE = EXP(LNLPP * LOG(2));
6         PROC MEANS MEAN;
7           VAR INTERCEP LNLPP EXPINTER SLOPE;
8         PROC PRINT;
9           VAR INTERCEP LNLPP EXINTER SLOPE;
```

Program IV - This program uses the first unit cost and slope coefficient data, in it's logarithmic state, from Programs I, II, and V. The PROC UNIVARIATE PLOT NORMAL statement is then used to provide information on the geometric mean of the first unit cost parameter, the range of the parameters sectioned by 1st-4th quartiles, and the maximum and minimum values of the parameters. Additional data such as normality plots, and the Kolomogorov-Smirnov D-statistic are also provided, but in these cases were not used.

```
1         LIBNAME SDAT '[TTRACHT];
2         DATA FIVE;
3         SET SDAT.PARAMS;
4         PROC SORT;
5           BY LNLPP;
6         PROC UNIVARIATE PLOT NORMAL;
7           VAR INTERCEP LNLPP;
```

Program V - This program performs WLSBF on the data generated in Programs II and III of Appendix A. The one difference between this program and Program II of Appendix B is the weighting scheme. Line 9 sets X equal to lot size and line 12 weights by X. The data regressed is the natural logarithm of either the heuristic lot plot point or the true lot midpoint, and average unit cost per lot. A permanent SAS data file is created which contains the intercept and slope coefficient data. A list file is also generated which contains ANOVA tables for each production run (100).

```
 1        LIBNAME MINE '[TTRACHT]';
 2        DATA THREE;
 3        INFILE ULOTGEN;
 4          INPUT CUMX TOTY LOTCOST LPP AVUNCST LOTSZ;
 5        INFILE PRODRUN;
 6          INPUT PRDRUN;
 7        LNLPP = LOG(LPP);
 8        LNAVUCST = LOG(AVUNCST);
 9        X = LOTSZ;
10        PROC REG OUTEST = MINE.WPARAMS;
11          MODEL LNAVUCST = LNLPP;
12            WEIGHT X;
13            BY PRDRUN;
14        PROC PRINT;
```

Program VI - This program uses the data generated in Program VI of Appendix A. The data is divided into four separate data sets by lot size. The PROC UNIVARIATE PLOT NORMAL statement provides detailed information on the distribution of the variables. In this case the variables are average unit cost per lot (X5) and the natural logarithm form of the same variable (LOGX5). This is accomplished for data grouped by lot sizes 1, 8, 64, and 512.

```
 1        DATA SIX;
 2        INFILE ULOTGEN3;
 3          INPUT X1 X2 X3 X4 X5 X6;
 4        LOGX5 = LOG(X5);
 5        DATA LOTSZ1;
 6          SET SIX;
 7          IF X6 = 1;
 8          PROC UNIVARIATE PLOT NORMAL;
 9            VAR X5 LOGX5;
10          PROC PRINT;
11        DATA LOTSZ8;
12          SET SIX;
13          IF X6 = 8;
14          PROC UNIVARIATE PLOT NORMAL;
15            VAR X5 LOGX5;
16          PROC PRINT;
17        DATA LOTSZ64;
18          SET SIX;
19          IF X6 = 64;
20          PROC UNIVARIATE PLOT NORMAL;
21            VAR X5 LOGX5;
22          PROC PRINT;
23        DATA LOTSZ512;
24          SET SIX;
25          IF X6 = 512;
26          PROC UNIVARIATE PLOT NORMAL;
27            VAR X5 LOGX5;
28          PROC PRINT;
```

Program VII - This program uses the data generated in Program VI of Appendix A. A new variable, the natural logarithm of the average unit cost per lot (LOGX5), is formed. The data is then sorted by lot size (X6), then using the Proc Means Var statement the variance of both X5 and LOGX5 are calculated. This provides the variance of both the standard and tranformed state of average unit cost per lot for lot sizes of 1, 8, 64, and 512.

```
 1          DATA SEVEN;
 2          INFILE ULOTGEN3;
 3            INPUT X1 X2 X3 X4 X5 X6;
 4          LOGX5 = LOG(X5);
 5          PROC SORT;
 6            BY X6;
 7          PROC MEANS VAR;
 8            VAR X5 LOGX5;
 9            BY X6;
10          PROC PRINT;
11            VAR X5 LOGX5;
```

## Biblography


1. Adams, Charles D. "The Improvement Curve Technique," Rockwell International Corporation, Reprinted November 1980.

2. Air Force Institute of Technology. Cost Improvement Analysis. A text on cost improvement curves for QMT180. Wright-Patterson AFB OH: School of Systems and Logistics, October 1986.

3. Air Force Systems Command. The AFSC Cost Estimating Handbook Series: Volume I "AFSC Cost Estimating Handbook". Reading MA: The Analytic Science Corporation, undated.

4. Aitchison, J. and J. A. C. Brown. The Lognormal Distribution. New York: The Syndics of the Cambridge University Press, 1957.

5. Avinger, Captain Charles R. Analysis of Learning Curve Fitting Techniques. MS Thesis. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1987.

6. Barry, John M. "Congress Wrestles the Pentagon on Procurement," Dun's Business Month, 126: 38-41 (August 1985).

7. Brewer, Glen M. The Learning Curve in the Airframe Industry. MS Thesis. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, August 1965.

8. Carrick, Paul M. "Experience Curves on Five Army Programs," Proceedings of the Department of Defense Annual Cost Analysis Symposium (17th). Arlington VA, September 1982.

9. Cochran, E. B. Planning Production Cost: Using the Improvement Curve. San Francisco: Chandler Publishing Company, 1968.

10. Daneman, Jeff. "Estimating Relationships With Log Linear Transforms," Unpublished course handout for Advanced Quantitative Methods for Cost Analysis QMT550. Wright-Patterson AFB OH: School of Systems and Logistics, undated.

11. -----. "How Good is our LSBF Equation at Predicting," Journal of Parametrics, 6: 52-60 (September 1986).

12. "Defence Procurement: A Job too Important for Servicemen," The Economist, 298: 31-34 (March 8, 1986).

13. Gansler, Jacques S. "Defense Program Instability: Causes, Costs, and Cures," Defense Management Journal, 22: 3-11 (Second Quarter 1986).

14. Guest P. G. Numerical Methods of Curve Fitting. Bristol, Great Britain: John Wright and Sons Ltd., 1961.

15. Hutchison, Capt Larry D. A Microcomputer Program for the Solution of Learning Curve Computations. MS Thesis. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1985 (AD-A161648).

16. Ilderton, R. B. Methods of Fitting Learning Curves to Lot Data Based on Assumptions and Techniques of Regression Analysis. MS Thesis. The Graduate School of Arts and Science, George Washington University, Washington DC, August 1970 (AD-A011 583).

17. Isaacson, Walter. "The Winds of Reform," Time, 121: 12-30 (March 7, 1983).

18. Johnson, J. Econometric Methods. New York: McGraw-Hill Book Company, 1984.

19. Kankey, Roland D. Assistant Professor of Quantitative Management. Personnal Interviews. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, May-August 1988.

20. Lieber, Raymond S. Production Cost Analysis Using the Underlying Learning Curve. MS Thesis. The Graduate School of Engineering, New Mexico State University, Las Cruces, New Mexico, April 1981.

21. Mendenhall, William. Introduction to Probability and Statistics. North Scituate, MA: Wadsworth Publishing Company, Inc., 1979.

22. Moskowitz, Herbert and Gordon P. Wright. Statistics for Management and Economics. Columbus OH: Charles E. Merril Company, 1985.

23. Murphy, Richard L. Assistant Professor of Quantitative Management. Personal Interviews. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, April-August 1988.

24. -----. Lecture notes from COST 672, Model Diagnostics and Software Management. School of Systems and Logistics, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, April 1988.

25. Nelson, Lt Cmdr Howard K. Ramifications of Learning Curve Theory and Use by Electronic Manufactures, MS Thesis. Naval Postgraduate School, Monterey CA, June 1985 (AD-B095541).

26. Neter, John, William Wasserman, and Michael H. Kutner. Applied Linear Statistical Models. Homewood, IL: Richard D. Irwin, Inc., 1985.

27. SAS User's Guide: Basics, Version 5 Edition. SAS Institute Inc., Cary NC, 1985.

28. SAS User's Guide: Statistics, Version 5 Edition. SAS Institute Inc., Cary NC, 1985.

29. Weisberg, Sanford. Applied Linear Regression. New York: John Wiley and Sons, Inc., 1985.

30. Wright, T. P. "Factors Affecting the Cost of Airplanes." Journal of the Aeronautical Sciences, 3: 122-128 (February 1936)

31. Yelle, Louis E. "The Learning Curve: Historical Review and Comprehensive Survey", Decision Science, 10: 302-328 (April 1979).

Vita

Captain Tom Tracht [redacted] Upon graduation he enlisted in the USAF. He was awarded a four-year ROTC scholarship through the Airman Education and Commissioning Program while stationed at Eielson AFB, Alaska. He graduated from the University of Washington in 1982 with a Bachelor of Science in Mathematics with a specialization in statistics. Upon graduation he received his commission in the USAF. After he was called to active duty in October 1982, he was employed as a program manager at the Deputy for Reconnaissance and Electronic Warfare, Aeronautical Systems Divison (ASD/RW). He then served as a cost analyst at the Engine System Program Office, Aeronautical Systems Division (ASD/YZ), until entering the School of Systems and Logistics, Air Force Institute of Technology, in 1987.

[redacted]

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>UNCLASSIFIED | 1b. RESTRICTIVE MARKINGS |
| 2a. SECURITY CLASSIFICATION AUTHORITY | 3. DISTRIBUTION/AVAILABILITY OF REPORT<br>Approved for public release; distribution unlimited. |
| 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>AFIT/GCA/LSQ/88S-9 | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>School of Systems and Logistics<br>6b. OFFICE SYMBOL (If applicable)<br>AFIT/LSQ | 7a. NAME OF MONITORING ORGANIZATION |
| 6c. ADDRESS (City, State, and ZIP Code)<br>Air Force Institute of Technology (AU)<br>Wright-Patterson AFB, Ohio 45433-6583 | 7b. ADDRESS (City, State, and ZIP Code) |
| 8a. NAME OF FUNDING/SPONSORING ORGANIZATION<br>8b. OFFICE SYMBOL (If applicable) | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER |
| 8c. ADDRESS (City, State, and ZIP Code) | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO |

11. TITLE (Include Security Classification)
AN ANALYSIS OF THE IMPACT OF LOG-LINEAR REGRESSION ON THE ESTIMATED LEARNING CURVE PARAMETERS

12. PERSONAL AUTHOR(S)
Tom Tracht, Captain USAF

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (Year, Month, Day) | 15. PAGE COUNT |
|---|---|---|---|
| MS Thesis | FROM _____ TO _____ | 1988 September | 114 |

16. SUPPLEMENTARY NOTATION

| 17. COSATI CODES | | |
|---|---|---|
| FIELD | GROUP | SUB-GROUP |
| 12 | 03 | |
| | | |

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Cost Estimates, Learning Curves, Least Squares Method, Curve Fitting

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Thesis Advisor: Roland D. Kankey
Assistant Professor of Quantitative Management

Approved for public release IAW AFR 190-1.

WILLIAM A. MAUER 17 Oct 88
Associate Dean
School of Systems and Logistics
Air Force Institute of Technology (AU)
Wright-Patterson AFB OH 45433

| | |
|---|---|
| 20. DISTRIBUTION/AVAILABILITY OF ABSTRACT<br>☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS | 21. ABSTRACT SECURITY CLASSIFICATION<br>UNCLASSIFIED |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>Roland D. Kankey | 22b. TELEPHONE (Include Area Code)<br>(513) 255-8409<br>22c. OFFICE SYMBOL<br>AFIT/LSQ |

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

BLOCK 19

ABSTRACT

This research had three main objectives. First, was to determine whether bias existed in the estimate of the learning curve parameters, as calculated by popular learning curve programs. Second, was to compare the fitting techniques of ordinary and weighted least-squares, with and without bias removed. Third, was to test the normality and constant variance assumptions of the average unit cost data.

When using the least-squares fitting technique to fit production lot data to the unit formulation of learning curve theory, log-linear regression biases the estimate of the first unit cost parameter on the high side, resulting in overestimating program costs. The bias was shown to be almost exclusively a function of the variance in the production cost data.

Factors which would reduce the bias were investigated. An approximation to the bias reduction factor proposed by Ilderton was used with excellent results.

The comparison of the ordinary and weighted least-squares fitting techniques produced mixed results. If the bias was not removed ordinary least-squares provided a better point estimate, although the dispersion was somewhat greater than with weighted least-squares. If the bias was removed the weighted least-squares technique proved to fit the data better in all areas.

Finally, the assumptions of normality and constant variance of the average unit cost data were tested. The results indicated that the assumptions were valid as long as the applicability of the model tested is accepted.